Escuela Politécnica Superior

Metodología ágil basada en KPI para la implantación de sistemas Big Data en empresas

Máster Universitario en Ingeniería Informática

Trabajo Fin de Máster

Autor:
Alberto Reales Díaz

Tutor:
Rafael Molina Carmona

Universitat d'Alacant
Universidad de Alicante

Enero 2019

Metodología ágil basada en KPI para la implantación de sistemas Big Data en empresas

Alberto Reales Díaz

Enero 2019

Universidad de Alicante
Máster Universitario en Ingeniería Informática
Tutor: Rafael Molina Carmona

Universitat d'Alacant
Universidad de Alicante

Agradecimientos

Este proyecto no hubiese sido posible sin la colaboración de Cruz Roja Alicante, por lo que el primer agradecimiento va para ellos, que se han volcado en su realización desde que se empezó a concebir, y han dado lo mejor de ellos de forma totalmente altruista, a pesar de lo apretado de sus agendas. Quiero mencionar, de forma particular, a Javier Rodríguez, responsable de Sistemas de Información de Cruz Roja, que ha dedicado tiempo, ilusión, y no ha dejado de aportar todo su conocimiento y experiencia a nuestro proyecto. Un millón de gracias, también, a Daniel Martínez, responsable de Relaciones Institucionales, que me ha brindado todo su apoyo en los momentos iniciales del proyecto, a Sandra Zambrana, técnico Provincial del Programa de Extrema Vulnerabilidad, que me ha enseñado las entrañas de la aplicación y me ha permitido conocer de cerca el funcionamiento de la misma, así como los procesos asociados y a Sagrario Sampere, directora del Plan de Intervención, que me ha explicado a la perfección la gran labor humana que realiza la organización y todo lo que aportan a la sociedad.

Otro gran apoyo durante todo el procedimiento ha sido mi tutor, Rafael Molina, del Departamento de Ciencias de la Computación e Inteligencia Artificial, de la UA. Desde que hicimos el proyecto en la asignatura ITA del Máster teníamos la idea de aprovechar el TFM para seguir colaborando con Cruz Roja. Su ayuda incondicional y su motivación han sido un acelerador que ha hecho de este un proyecto mejor. Todos sus consejos, observaciones, tirones de oreja y felicitaciones han hecho que este trabajo sea posible.

Por último, pero no por ello menos importante, quiero dedicar este proyecto a mi familia, que siempre ha creído en mi, a mis amigos, que siempre tienen una cerveza fría cuando es necesario, y a mis compañeros de trabajo, que no dejan de hacerme sugerencias para seguir mejorando día a día. Este trabajo os lo dedico a vosotros Milagros, Enrique, Matías, Toni, Mercedes, Clara, Patricia... Imposible recoger todos los nombres en un papel de espacio finito. ¡Gracias a todos y cada uno de vosotros por estar a mi lado!

Resumen

El presente proyecto aborda el diseño de una nueva metodología que pretende resolver una serie de problemas muy comunes tanto en organizaciones como en empresas o administraciones públicas. Necesitan realizar proyectos para generar conocimiento a partir de sus datos, pero no cuentan con los perfiles profesionales, las infraestructuras o la estructura directiva necesarias, por lo que deben adaptar todo su negocio y adquirir nuevo conocimiento antes y durante la realización de dichos proyectos. Por este motivo, es muy habitual que estos proyectos empiecen cubriendo áreas muy concretas y pequeñas de la compañía a modo de prueba de concepto para que el equipo pueda empezar a formarse, y luego, ir desarrollando proyectos mayores en áreas más estratégicas.

La metodología propuesta en este proyecto pretende cubrir este proceso de aprendizaje. Las metodologías existentes, como MBDA, están pensadas para proyectos de cierta envergadura, donde el equipo es lo suficientemente experimentado como para definir los parámetros del proyecto desde el principio. En nuestro caso, pretendemos que el equipo defina objetivos base que sirvan para ganar experiencia y construir, de forma iterativa, sistemas cada vez más complejos y útiles para la compañía, añadiendo valor durante cada paso del proceso.

Se ha seguido una aproximación iterativa para desarrollar el proyecto, partiendo de un estudio de las metodologías previas mencionadas que son modificadas para adaptarse a los objetivos que se pretenden cubrir. Como banco de pruebas, hemos utilizado un contexto real dentro de la organización Cruz Roja, que ha colaborado con este proyecto para que podamos realizar una iteración completa de la metodología tal cual se hubiese aplicado en el contexto real. Gracias a esta colaboración, ha sido posible comprobar hasta qué punto el diseño se adapta a un entorno real para ir modificándolo hasta que el resultado ha sido el deseado.

Los resultados han sido satisfactorios. Se ha obtenido una metodología que tiene en cuenta las restricciones del dominio para la que está pensada y que puede ser un excelente punto de partida para elaborar un estándar para el desarrollo de proyectos Big Data.

Sin duda, es muy necesario que se investigue en este sentido y se trate de elaborar estándares que unifiquen los criterios de despliegue y calidad de los sistemas Big Data. Dado los resultados obtenidos, se puede decir que este proyecto es un paso en ese sentido.

Resum

El present projecte aborda el disseny d'una nova metodologia que pretén resoldre una sèrie de problemes molt comuns tant en organitzacions com en empreses o administracions públiques. Necessiten realitzar projectes per generar coneixement a partir de les seues dades, però no compten amb els perfils professionals, les infraestructures o l'estructura directiva necessàries, de manera que han d'adaptar tot el seu negoci i adquirir nou coneixement abans i durant la realització d'aquests projectes. Per aquest motiu, és molt habitual que aquests projectes comencen cobrint àrees molt concretes i reduïdes de la companyia a manera de prova de concepte perquè l'equip puga començar a formar-se, i després, anar desenvolupant projectes majors en àrees més estratègiques.

La metodologia proposada en aquest projecte pretén cobrir aquest procés d'aprenentatge. Les metodologies existents, com MBDA, estan pensades per a projectes de certa envergadura, on l'equip és prou experimentat com per definir els paràmetres del projecte des del principi. En el nostre cas, pretenem que l'equip definixca objectius base que servixquen per guanyar experiència i construir, de manera iterativa, sistemes cada vegada més complexos i útils per a la companyia, afegint valor durant cada pas del procés.

S'ha seguit una aproximació iterativa per desenvolupar el projecte, partint d'un estudi de les metodologies prèvies esmentades que són modificades per adaptar-se als objectius que es pretenen cobrir. Com a banc de proves, hem utilitzat un context real dins de l'organització Creu Roja, que ha col·laborat amb aquest projecte perquè puguem fer una iteració completa de la metodologia tal qual s'haguera aplicat en el context real. Gràcies a aquesta col·laboració, ha sigut possible comprovar fins a quin punt el disseny s'adapta a un entorn real per anar modificant fins que el resultat ha sigut el desitjat.

Els resultats han sigut satisfactoris. S'ha obtingut una metodologia que té en compte les restriccions del domini per a la qual està pensada i que pot ser un excel·lent punt de partida per elaborar un estàndard per al desenvolupament de projectes Big Data. Sens dubte, és molt necessari que s'investigue en aquest sentit i es tracte d'elaborar estàndards que unifiquen els criteris de desplegament i qualitat dels sistemes Big Data. Donat els resultats obtinguts, es pot dir que aquest projecte és un pas en aquest sentit.

Índice de contenidos

Índice de figuras

Índice de tablas

1. Introducción

Cada vez es más frecuente que las empresas emprendan proyectos destinados a aprovechar la gran cantidad de datos con la que cuentan para poder tomar mejores decisiones y ser más competitivos en los mercados actuales. Es muy habitual que las compañías tengan grandes cantidades de datos derivados de sus procesos que no son aprovechados, bien porque no se cuenta con la tecnología necesaria, o bien porque no se ha valorado el impacto económico y productivo que pueden tener dichos datos en la compañía. Del mismo modo, hay casos de éxito en empresas que han hecho grandes esfuerzos por extraer valor de estos datos y han conseguido resultados excelentes [32]. Esto ha impulsado el interés de la mayoría de empresas por el Big Data abriendo un nuevo horizonte al ofrecer la posibilidad de obtener mucho más conocimiento de la información disponible en la organización de lo que ha sido posible hasta ahora.

De este modo, surge la necesidad de establecer metodologías y procedimientos que permiten llevar a cabo este tipo de proyectos teniendo en cuenta sus particularidades. Un proyecto de análisis de datos masivos, típicamente conocidos como Big Data, presenta unas dificultades que no son contempladas en las metodologías de desarrollo software tradicionales. Lo que hace que no puedan ser utilizadas para proyectos Big Data es que son demasiado rígidas y están demasiado enfocadas a definir las arquitecturas software para cumplir los requisitos, y no tanto a que el negocio pueda aprovechar los datos y la información disponible. Además, los proyectos Big Data son mucho más cambiantes debido a la falta de madurez de este paradigma, lo que descarta metodologías que exijan definir, desde un primer momento, todos los requisitos del proyecto. Hasta ahora, en cada contexto se ha seguido una metodología propia por la falta de estándares y documentación [27].

En los proyectos Big Data se deben combinar muchas metodologías para capturar los requerimientos del negocio y que el resultado esté alineado con la estrategia de la compañía. Por tanto, se debe definir una sólida capa de software que sea capaz de escalar y cambiar junto con las necesidades del negocio, aplicarse los más modernos métodos de despliegue de aplicaciones para reducir el *time-to-market* y ser lo más competitivo posible. Además, se deben utilizar los modelos matemáticos necesarios y óptimos que permitan resolver el problema, todo ello sobre una infraestructura que debe recibir grandes cantidades de datos desde diferentes fuentes y con mucha más frecuencia de lo que son capaces de soportar las arquitecturas hardware tradicionales. Se podría decir que estos proyectos ponen a prueba tres perfiles profesionales: El científico de datos, encargado de capturar los requisitos y definir un modelo matemático que cubra dichos requisitos, así como una presentación de resultados adecuada; el ingeniero de datos, que se encarga de diseñar la aplicación, conectar las fuentes de datos y organizar los despliegues para que se puedan cumplir las exigencias del negocio; y el ingeniero de sistemas que debe asegurarse de que la infraestructura pueda dar un soporte adecuado a la capa software.

En este trabajo de final de máster se pretende abordar el diseño de una metodología que tenga en cuenta todos estos factores, conjugando lo mejor de las metodologías existentes para facilitar la implantación de proyectos Big Data a empresas tradicionales. Además, para validar la metodología, contaremos con la colaboración de una organización real, Cruz Roja, que nos servirá como banco de pruebas para refactorizar nuestro método y ajustarlo perfectamente al mundo real.

Analizar la inmensa cantidad de datos con la que cuenta Cruz Roja supone todo un

reto. Es un claro ejemplo de organización que surgió mucho antes de la era digital, y que ha ido adoptando las tecnologías de la información de forma progresiva, creando pequeñas aplicaciones que van resolviendo problemas concretos. Como consecuencia de estos procesos, existen multitud de fuentes de datos diferentes y aplicaciones que producen y consumen dichos datos, lo que hace que el entorno en el que se desarrolla el proyecto sea realmente complejo. Algunos de los problemas típicos que podemos encontrar en este tipo de entorno son [13]:

- Volumen: muchas aplicaciones acumulando datos durante muchos años generan un volumen que es condicionante para la arquitectura del sistema que pretenda realizar cualquier tipo de análisis sobre ellos.
- Variedad: el hecho de que existan multitud de fuentes de datos aisladas implica que la tecnología tendrá que ser capaz de conectarse a esas fuentes de datos, que normalmente serán heterogéneas (bases de datos relacionales, ficheros en disco de diferentes formatos, API's, etc.), e integrar toda la información en un único lugar para poder realizar operaciones sobre ella.
- Ruido en los datos: es muy habitual encontrar bases de datos que han almacenado datos sin ningún tipo de filtro de validación previa, lo que suele provocar que encontremos inconsistencias como datos faltantes, datos erróneos (poblaciones que no existen, códigos postales incompletos, etc.) o datos que no son coherentes entre ellos como, por ejemplo, que dos datos reflejen dos características incompatibles de un individuo, como indicar un código postal que no se corresponde con la ciudad informada en otro campo.
- Datos caducos: hay datos que es necesario actualizar o descartar ya que la probabilidad de que reflejen la realidad desciende conforme avanza el tiempo. Por ejemplo, en el caso del género, es un dato muy consistente en el tiempo debido a que es muy poco probable que cambie. Por otro lado, el estado civil de un individuo debe actualizarse cada poco tiempo ya que es algo mucho más cambiante para la población en general.
- Redundancia: Cuando hay tantas aplicaciones que han sido desarrolladas de forma paralela e independiente es muy habitual que se guarde el mismo dato de forma duplicada, haciendo que sea redundante e incluso inconsistente (el mismo dato puede estar informado de forma diferente en dos bases de datos distintas). Para corregir este problema, es necesaria una buena estructuración de los datos, que pasa por un buen diseño de la capa de persistencia de las aplicaciones.
- Resistencia al cambio: es habitual encontrarse con profesionales acostumbrados a los procesos manuales y reticentes a la automatización, lo que condiciona nuestra metodología para contemplar el proceso de cambio de mentalidad, mostrando el potencial de las nuevas tecnologías.

Veremos cómo las nuevas tecnologías y arquitecturas permiten poner en marcha estos proyectos [3, 5], garantizando la escalabilidad, la potencia y la resiliencia necesarias para el éxito, y cómo nuestra metodología consigue solventar los problemas que platean las metodologías tradicionales en este tipo de proyectos y empresas.

2. Marco teórico

2.1. Metodologías utilizadas en proyectos similares

Es muy importante comprender qué metodologías se han utilizado en proyectos similares para analizar sus fortalezas y debilidades. De este modo, podremos crear un método adaptado a este tipo de problemas y que pueda servir de estándar en un futuro. Nos centraremos, sobre todo, en la necesidad de integrar a los responsables del negocio en todo el desarrollo del proyecto, en garantizar el anonimato de los datos y en el que el proyecto resuelva los problemas de la organización [6] por ser factores claves para los proyectos Big Data, mucho más integrados en el negocio, si cabe, que los proyectos software tradicionales.

También nos apoyamos en metodologías de diseño software en general y organización de equipos ya que este problema tiene muchos puntos en común que han sido estudiados en dichas metodologías. Vamos a mencionar algunas de ellas para mostrar dichos elementos comunes:

- Test Driven Development (TDD) [29]: esta metodología trata de establecer los escenarios descritos por el usuario antes de empezar la codificación, garantizando que solo se implementa lo que se necesita, es decir, lo que se corresponde con los escenarios. Este es un punto clave de nuestra metodología, que trata de averiguar qué escenarios debe resolver mediante la creación de los KPI, que reflejan estos escenarios.

- Ingeniería de requisitos [23, 24]: los estándar creados para captar de un modo fiel los requisitos marcados por los usuarios tratan de asegurarse de que se obtenga un software que cubra todas las necesidades del usuario manteniendo un compromiso de coste económico y temporal. En nuestra metodología se tienen estos factores en cuenta mediante la priorización de los KPI que deben ser desarrollados y la integración del usuario en el desarrollo de la aplicación.

- SCRUM [31]: uno de los objetivos que persigue esta metodología de desarrollo pretende involucrar al usuario en el desarrollo software para que pueda seguir su evolución y aportar valor. Esto, combinado con que SCRUM divide el desarrollo en *sprints* o iteraciones, permite que el producto software resultante sea mucho más fiel a lo que el usuario necesita que en los desarrollos tradicionales en cascada. En nuestra metodología pretendemos exactamente esto, que el usuario aporte valor a todo el proceso y que participe de la información y de las conclusiones que se van alcanzando durante el proyecto, tratando de asegurar que la solución Big Data aporta todo el valor posible al negocio.

Una vez analizadas las metodologías software en las que se apoya nuestra metodología ágil, vamos a exponer dos buenos ejemplos de metodologías utilizadas en proyectos similares, analizando sus fortalezas, debilidades y puntos de referencia para la metodología propuesta en este proyecto.

2.1.1. Model-based Big Data Analytics-as-a-Service (MBDAaaS)

A pesar de no ser una metodología estándar, MBDAaaS trata de generalizar una metodología tipo MBDA, ampliamente utilizada en proyectos Big Data, ofreciéndola

con el paradigma *as a service* para que pueda ser utilizada en contextos generales. Por este motivo, supone una buena referencia, ya que nuestra metodología trata de ser lo suficientemente general como para ser aplicada en el contexto de cualquier organización.

MBDAaaS [6] trata de salvar la diferencia que existe entre las capacidades técnicas necesarias para poder desarrollar un proyecto de Big Data y el conocimiento de las necesidades de negocio. Esta metodología se basa en un modelo declarativo, en el que los responsables de negocio establecen los objetivos que quiere conseguir el proyecto en el formato "Indicadores/Objetivos", que son priorizados por el propio cliente. A partir de aquí, los indicadores y los objetivos son utilizados para establecer requisitos, tanto funcionales como no funcionales, del sistema Big Data. Todo esto deriva en la creación de un modelo procedural (Sección IV-C de la referencia) y un modelo de despliegue (Sección IV-D de la referencia).

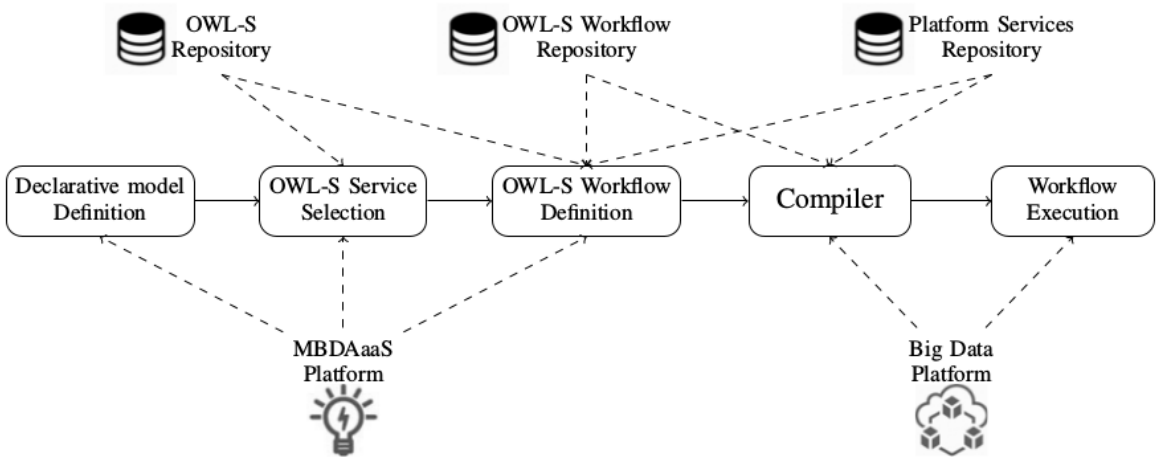


Figura 1: Secuencias de ejecución definidas por la metodología MBDAaaS [6]

Como puede apreciarse en la figura 1, esta metodología utiliza dos plataformas: la plataforma MBDAaaS que define la propia metodología y que recoge la información necesaria para llevar a cabo el proyecto, y la plataforma Big Data que se encarga de generar el código y hacer el procesamiento en paralelo de los datos. En nuestro caso, no necesitamos una plataforma que recoja esta información, sino que aprovechamos herramientas de comunicación gratuitas, ágiles y flexibles, como Trello[1]. Estas herramientas son de fácil acceso, con una curva de aprendizaje muy suave para usuarios que no tienen poco experimentados con las tecnologías y permiten colaborar para cambiar los parámetros del proyecto. Además, están pensadas para proyectos de pequeño o mediano tamaño. Las metodologías analizadas en nuestro contexto están pensadas para plantear proyectos de gran envergadura, añadiendo demasiada complejidad cuando el trabajo que se quiere desarrollar es más sencillo.

De esta metodología es muy útil aprovechar el modo de establecer los requisitos mediante los pares «Indicadores/Objetivos», que es algo muy natural para los clientes, ayudando a salvar la distancia que existe entre el cliente final y los analistas. No obstante, es demasiado lineal, y pierde la capacidad de adaptación necesaria en aquellos proyectos en los que no se conoce el proceso que se debe seguir, pero sí la meta que se quiere alcanzar.

[1]https://trello.com/

Nos encontramos en un escenario en el que el usuario, por falta de experiencia, inicialmente no es capaz de aprovechar todo el potencial de la tecnología. Este hecho hace que la metodología necesite tener una gran capacidad de adaptación, ya que los objetivos se irán refactorizando conforme el proyecto avance. El motivo de este desconocimiento es que se cuenta con tantos datos que es muy difícil estimar todo su potencial, que solo queda revelado cuando se empiezan a analizar los datos y a conseguir resultados parciales.

2.1.2. Metodología Iterativa

Hemos escogido la *Metodología Iterativa* como referencia porque intenta proponer una solución ágil a la definición de proyectos basados en datos, objetivo compartido por nuestra propia metodología.

Esta se basa en cinco fases fundamentales, de las cuales en este proyecto destacaremos la primera [33], ya que se centra en cómo llevar a cabo la fase de análisis del proyecto, realizada por el analista y el arquitecto de datos.

Esta metodología, en su primera fase, define cuatro subprocesos, que son:

1. Identificar las tareas de análisis que habrá que desarrollar.
2. Definir los puntos clave de los requerimientos no funcionales para cada tarea de análisis.
3. Agrupar las tareas de análisis que tienen los mismos valores para, después, gestionar cada uno de estos grupos como una sola tarea de análisis.
4. Definir el plan que establece qué fases se llevarán a cabo con los datos. En cada uno de los grupos, se establece cómo los datos evolucionarán.

Al ser una metodología iterativa, muy parecida a la propuesta en este proyecto, nos sirve como un punto de referencia. Además de estar planteada para el mismo tipo de implantaciones en entornos reales, no académicos, lo que la hace más adecuada. No obstante, como muestra la figura 2, se trata de una metodología muy compleja, pensada para grandes proyectos. Por eso, nuestro diseño resulta una simplificación del actual, haciéndolo más adecuado para equipos pequeños o proyectos iniciales en compañías tradicionales.

Hemos escogido estas dos metodologías porque son representativas en cuanto al tipo de métodos que se siguen en los proyectos Big Data, aunque ninguna se ha tomado como estándar de momento.

2.2. Tecnologías Big Data

En este punto presentamos una breve descripción de las tecnologías más empleadas en los proyectos Big Data. Todas ellas tienen un punto en común, la escalabilidad, que se consigue, de forma habitual, con el uso de arquitecturas tipo clúster. Concretamente, estamos hablando de escalabilidad horizontal, que es la que consiste en añadir más nodos en paralelo a la infraestructura para aumentar la potencia de cálculo sin afectar a los procedimientos existentes. La escalabilidad contraria, la vertical, consistente en aumentar los recursos con los que cuenta cada nodo, no es el objetivo de este tipo de soluciones debido a que requiere que sean reiniciados los nodos, lo que no es operativo

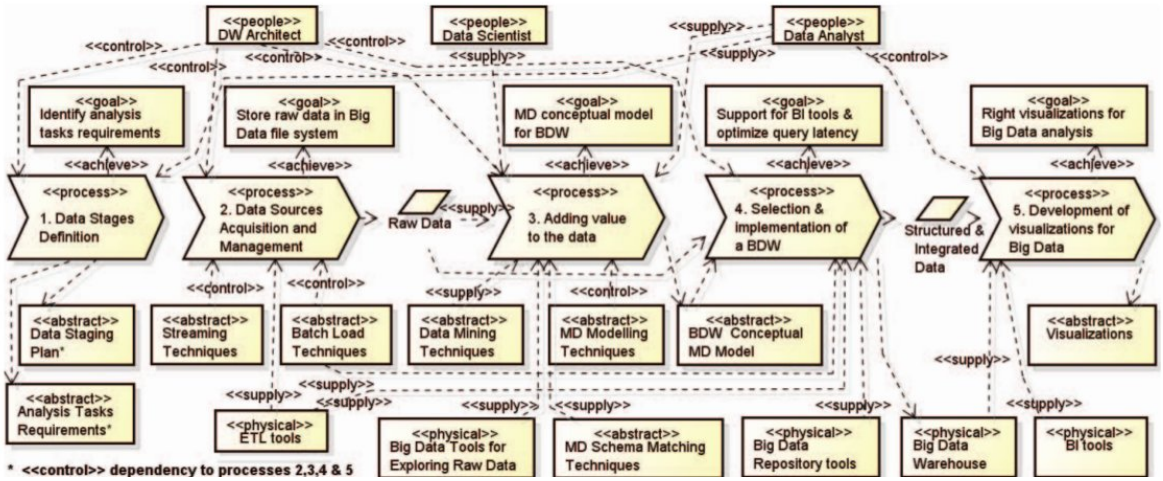


Figura 2: Diagrama de la metodología. [33]

en entornos reales. Actualmente, esta característica es vital para cualquier sistema que pretenda abarcar un proyecto real en este ámbito, ya que si el sistema no es capaz de escalar de forma sencilla y transparente, quedará obsoleto rápidamente, y su actualización provocará cortes en el servicio que, a su vez, provocarán retrasos en los resultados.

A continuación, enumeramos estas tecnologías, con una breve descripción de ellas.

2.2.1. Apache Hadoop

Apache™Hadoop es una librería de código libre que permite realizar procesamiento de datos distribuido en un clúster de computadores utilizando modelos de programación sencillos [3]. Esta tecnología es la base de todas las demás, como puede verse en la figura 3, ya que facilita la gestión de los recursos del clúster automatizando la detección de fallos, la incorporación y eliminación de recursos, un sistema de almacenamiento y computación distribuido, etc.

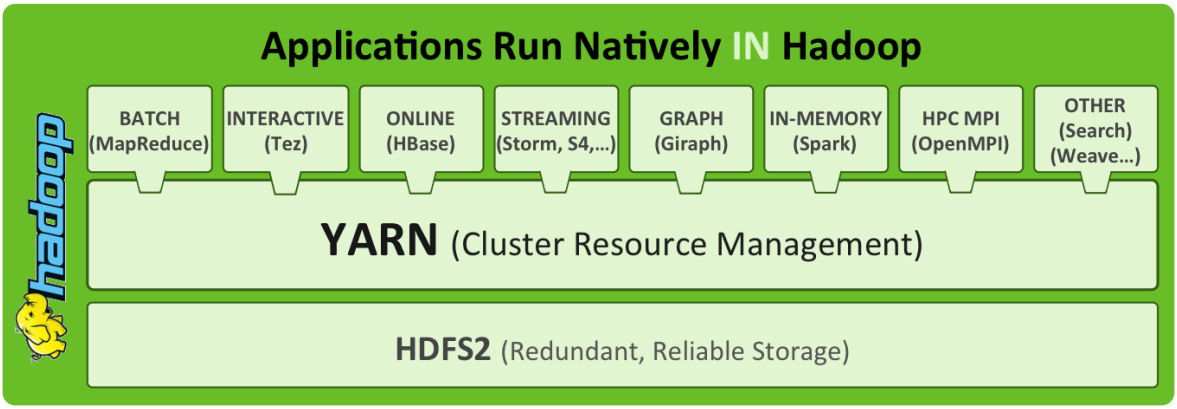


Figura 3: Stack de Hadoop (Fuente: Página oficial de Hadoop)

Algunos de los módulos más relevantes que incluye este proyecto son:

- **Hadoop Distributed File System (HDFS™)**: es el sistema de almacenamiento distribuido que utiliza Hadoop, y proporciona las siguientes ventajas:

  - Escalabilidad: cuando se añade un nuevo nodo al clúster, este pasa a formar parte del HDFS. Automáticamente, Hadoop calculará qué datos debe almacenar y los enviará al nuevo nodo, reconfigurando el almacenamiento del clúster de forma automática.
  - Resiliencia: cada dato es troceado en bloques, que son almacenados de forma redundante entre los nodos del clúster de modo que, en total, hay varias copias del mismo bloque en el clúster. Esto garantiza que, si cae un nodo, el dato esté disponible en otro nodo. En este caso, si existen menos de N copias en el clúster, HDFS creará una nueva copia en otro nodo para garantizar este funcionamiento.
  - *Data Locality*: esta propiedad consiste en realizar una determinada operación en el mismo nodo en el que se encuentra el dato que necesita, con el objetivo de minimizar el tráfico de red y, por lo tanto, el tiempo total de cómputo. Gracias a que HDFS guarda varias copias del mismo dato, es posible seguir este tipo de estrategias. Lo mejor de todo es que Hadoop gestiona todo esto de forma transparente, sin que el programador tenga que preocuparse por ello.

- **Hadoop Yarn**: este framework provee una mejor gestión de los recursos del clúster que su predecesor, MapReduce, a la hora de ejecutar los jobs programados [28, 8]. Permite ejecutar de forma muy eficiente los jobs, ya que planifica qué ejecutores deben llevar a cabo las operaciones teniendo en cuenta el principio de *data locality*, la carga de trabajo de los nodos e, incluso, el tiempo que tardan en responder los mismos. Si un nodo se retrasa en procesar una tarea, automáticamente, y sin esperar a que este reporte un fallo, lanzará esa misma tarea en otro nodo, dando por válido el primer resultado que se obtenga. Esto nos da una idea del nivel de automatización y optimización de este framework.

2.2.2. Apache Spark

Apache Spark es una plataforma de computación distribuida optimizada para conseguir rendimientos adecuados para el procesamiento de datos en tiempo real [14]. Esta optimización se consigue operando con los datos en memoria RAM. Por supuesto, esto requiere que el clúster cuente con gran cantidad de memoria, lo que condiciona nuestra inversión en infraestructura.

Como puede verse en la siguiente figura, donde se muestra el tiempo, en segundos, que tarda una regresión logística en calcularse sobre Hadoop mediante *map reduce* y sobre Spark, el rendimiento aumenta de forma considerable, lo que justifica la inversión en un clúster adecuado.

Sin embargo, lo más interesante de Apache Spark es la cantidad de librerías con las que cuenta, que van desde *Machine Learning*, como Spark MLlib hasta librerías de visualización, como GraphX. Estas librerías otorgan un buen grado de generalidad a Spark, por lo que podremos adaptarlo fácilmente a la resolución de problemas que queramos resolver en nuestro negocio.

En la figura 5 (pág 22) podemos encontrar el stack de tecnologías que se basan en Spark.

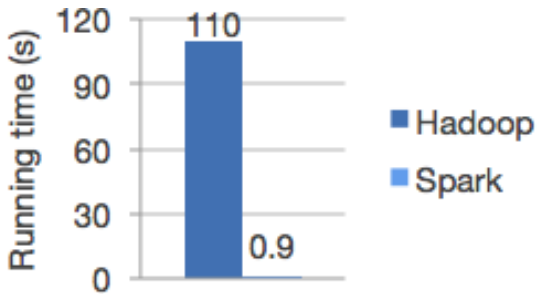


Figura 4: Tiempo, en segundos, de cálculo de una regresión logística en Hadoop vs Spark (Fuente: Web oficial de Spark)

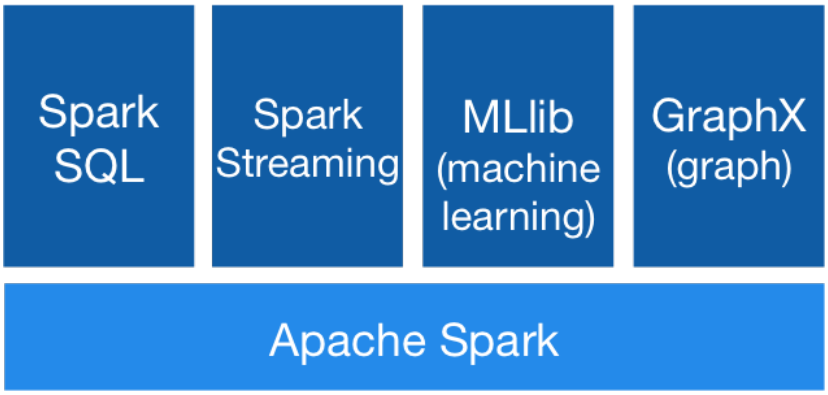


Figura 5: Stack de librerías de Spark (Fuente: Web oficial de Spark)

2.2.3. Apache Parquet

Apaque Parquet [4] es un formato de almacenamiento de datos en forma de columna, el cual consigue un gran rendimiento gracias a su optimizada estructura y a la compresión conseguida.

El rendimiento conseguido por utilizar este formato a la hora de almacenar los datos justifica su uso en cualquier nuevo proyecto. Además, es compatible con cualquier tecnología Big Data como Spark, Hive o el propio Hadoop, y puede ser creado a partir de otros formatos, como los populares CSV y JSON, convirtiéndose, así, en un punto intermedio ideal entre diferentes fuentes de datos, ya sean las mencionadas o fuentes de datos en tiempo real como, por ejemplo, un log generado con Apache Kafka[2].

2.3. Cloud computing

Si hay algo que todos los expertos en Big Data tienen claro es la revolución que ha supuesto para este paradigma la irrupción de las tecnologías cloud, que han hecho posible crear y gestionar estructuras realmente complejas mediante sencillas interfaces gráficas y paneles de configuración. Hasta su aparición, la única opción era crear un cluster de Hadoop en los propios servidores de la compañía, delegando, a menudo, la responsabilidad de su creación, configuración y gestión a los equipos de infraestructura tradicionales, que debían reciclarse para obtener el conocimiento necesario para su gestión.

Las plataformas como Cloudera [10] o Hortonworks [11] facilitaron mucho esta labor,

[2]https://kafka.apache.org

pero todavía tenían una carencia, y es que seguían dependiendo de la infraestructura propia de la compañía.

Esta suele ser poco flexible. Normalmente, se asignan una serie de recursos a un proyecto, en nuestro caso, al *cluster* destinado a los proyectos Big Data, y se deja habilitado. En los momentos de mayor necesidad, la infraestructura se queda corta, ralentizando la obtención de resultados. Cuando no se están lanzando trabajos contra el cluster, las máquinas no se apagan, por lo que todos esos recursos están siendo desaprovechados. Para conseguir esta elasticidad, se podría recurrir a tecnologías como Kubernetes, que podrían levantar máquinas virtuales de forma dinámica utilizando contenedores, pero los servidores físicos seguirían siendo un problema.

Las plataformas cloud resuelven ambos problemas, ya que permiten crear, en segundos, clusters con las capacidades necesarias, lanzar un análisis, persistir la información y los resultados necesarios, y desconectar el cluster, ahorrando recursos y garantizando un tiempo de respuesta muy rápido. Todo ello, con una fácil gestión.

Además, estas plataformas ofrecen soluciones integrales para el Big Data, por lo que también cuentan con herramientas de visualización de datos, diferentes tipos de almacenamiento, integraciones con plataformas IoT (Internet of Things), herramientas de consulta de datos como Hive, etc.

2.4. Técnicas de análisis de datos

Existen multitud de técnicas que nos ayudan a analizar el estado en el que se encuentran nuestros datos para sacar conclusiones previas y optimizar nuestro dataset para conseguir que nuestro modelo obtenga mejores resultados. A continuación, vamos a hacer una breve explicación de algunas de ellas.

2.4.1. Reducción de dimensionalidad

Todas las técnicas estadísticas descritas en este apartado tienen como objetivo reducir la dimensionalidad de nuestro dataset sin sacrificar calidad en los datos. Los beneficios de aplicar este tipo de técnicas a nuestro conjunto de datos antes de procesarlos son reducir el tamaño del dataset, reducir su complejidad, lo que provoca que los modelos que se obtenga sean más sencillos también, y detectar redundancias que pueden hacer que se replanteen los datos que se están recogiendo desde negocio, eliminando estos datos repetidos o estrechamente relacionados.

Análisis de componentes principales El análisis de componentes principales detecta columnas de nuestro dataset que son iguales o están íntimamente relacionadas, de forma que permiten obtener exactamente la misma información, sin aportar nada nuevo a los modelos de *machine learning*. Estas columnas son eliminadas dejando en el dataset una o varias nuevas componentes que estén correlacionadas, de forma directa o indirecta, con las originales, conservando la misma información en menos dimensiones y, por lo tanto, en menor tamaño.

3. Objetivos

3.1. Objetivo general

En el presente proyecto pretendemos conseguir una metodología que permita implantar, de forma ágil, proyectos de Big Data en empresas con pocos recursos y/o poca experiencia en proyectos similares.

3.2. Objetivos específicos

En este proyecto se persiguen los siguientes objetivos secundarios:

- Estudiar metodologías existentes para proyectos Big Data académicos o empresariales, analizando sus fortalezas y debilidades.
- Proponer una nueva metodología inspirada en las fortalezas de los modelos anteriores, que mejore sus debilidades mediante su adaptación al caso de empresas que abordan su primer proyecto de este tipo.
- Validar la metodología aplicándola, al menos en parte, en un entorno real.

3.3. Acciones requeridas

Para conseguir los objetivos propuestos en el proyecto, son necesarias las siguientes acciones:

- Documentar metodologías existentes.
- Entrevistar a los responsables de TI y de negocio de una empresa real.
- Documentar el flujo de datos de dicha empresa.
- Realizar una iteración completa de la metodología.

4. Metodología para el desarrollo del proyecto

En este capítulo explicaremos el método seguido para la realización del presente trabajo final de máster. Es importante no confundir este método, que es el escogido para la elaboración del trabajo, con el resultado del propio trabajo, que es una metodología para implantación de proyectos Big Data. Para elaborar este proyecto, se ha seguido una metodología basada en la ingeniería de requisitos tradicional, en la que se enfoca el problema desde las necesidades del negocio hacia la solución técnica, y no al revés [25]. La captación de requisitos se ha hecho de forma incremental e involucrando a la capa de negocio de Cruz Roja para asegurarnos de que el resultado se ajusta a las necesidades de una organización real. De modo que los pasos seguidos para la elaboración del proyecto han sido:

1. Primera reunión con los responsables de tecnología de Cruz Roja, en la que se explica el alcance que se pretende abordar con el proyecto para que se puedan establecer una o varias áreas candidatas para desarrollar el proyecto.

2. Entrevistas con los responsables de cada una de las áreas establecidas en el paso anterior para valorar el impacto de la integración respectivamente, solicitando documentación que permita conocer y estudiar cada escenario particular de forma detenida.

3. Una vez estudiada la documentación, se elabora una propuesta de proyecto preliminar para cada área que nos sirva como punto de partida para escoger la de actuación final, estableciendo el alcance final del proyecto.

4. Reunión con el responsable del área en la que se documentan y priorizan las necesidades del negocio para establecer las fases del proyecto.

5. Con los requisitos de la parte del negocio claros, comenzamos a estudiar las metodologías existentes para llevar a cabo este tipo de proyectos, estudiando sus fortalezas y debilidades, que nos servirán para elaborar nuestra propia metodología.

6. Se comienza el diseño de la metodología mediante iteraciones. En cada iteración, se valida con la capa de negocio la viabilidad de los cambios introducidos. De este modo, nos aseguramos de que se podrá aplicar sin problemas en un entorno real, teniendo en cuenta las restricciones que nos vamos encontrando.

7. Una vez finalizado el diseño de la metodología, pasamos a la fase de implementación, en la que realizaremos el despliegue del proyecto durante una iteración para validar la metodología. Durante esta fase, todavía es posible refactorizar la metodología final.

8. Finalización del proyecto y elaboración de conclusiones.

A modo de conclusión de este apartado, podemos decir que, siguiendo esta metodología, hemos conseguido:

- Elaborar la metodología de realización de proyectos Big Data en organizaciones o empresas de pequeño tamaño y/o poca experiencia en el ámbito del Big Data.

- La validación de dicha metodología mediante su aplicación en un caso real.
- La propia memoria del trabajo final de máster, donde se recoge todo el trabajo realizado.

5. Metodología ágil basada en KPI para la implantación de sistemas Big Data en empresas

En este trabajo se presenta una metodología iterativa cuyo objetivo es abordar el análisis de los datos de la compañía y el despliegue de las aplicaciones que los explotan por fases, de forma que podamos refinar la arquitectura y los modelos de datos en base a los resultados obtenidos en cada iteración para ir progresando hacia el resultado deseado.

Es necesario definir una metodología adecuada a este tipo de proyectos dado que no existen estándares para realizar el análisis que pretende llevarse a cabo [33, 13], sino que se crean métodos acoplados a cada escenario concreto, lo que obliga al especialista en datos a diseñar metodologías específicas que se adapten al entorno en el que se desarrolla el proyecto basándose en casos de éxito anteriores. En este punto, la capacidad de adaptación y la experiencia del analista son fundamentales para el éxito del proyecto. De este modo, si el especialista cuenta con un estándar que le sirva de guía a la hora de desarrollar el proyecto de análisis de datos, podrá obtener la información necesaria de forma mucho más efectiva, mejorando el resultado final.

Con esta metodología se pretende ofrecer una guía para aquellas empresas que pretenden comenzar con el análisis de datos de su negocio de menos a más, involucrándose en el proceso para fortalecerlo y enriquecerlo. Así, pretendemos que sea lo más sencillo posible realizar proyectos de Big Data en empresas tradicionales, que deben ir adaptando sus procesos de negocio y sus sistemas a este nuevo paradigma afectando lo menos posible a su día a día. Este es el escenario de muchas empresas que se encuentran con que son incapaces de abordar un proyecto de esta magnitud por sí mismas y que, incluso, se ven perdidas ante la falta de estándares y documentación por lo reciente que es la aparición de estas tecnologías a nivel de empresas.

5.1. Factores condicionantes de la metodología

Para poder diseñar una metodología que se adapte a una empresa real, es necesario que se tengan en cuenta los factores que condicionan nuestro diseño. En esta sección, solo vamos a tener en cuenta los factores que afectan, en general, a este tipo de proyectos, pero a la hora de aplicarla será necesario valorar las particularidades del escenario en el que nos encontremos para poder seguir esta metodología de la forma correcta y que el resultado sea óptimo.

El principal factor condicionante es la necesidad de integrar al equipo de profesionales existente en el área que se verá afectada por el despliegue del proyecto. De este modo, estarán involucrados en el proceso de descubrimiento de información y podrán confirmar si las conclusiones que se están obteniendo son útiles para ellos y reconducir el trabajo de las siguientes iteraciones si lo consideran necesario. Las metodologías tipo TDD (Test-driven development) o BDD (Behavioural-driven development) son una excelente referencia ya que parten desde el usuario, teniendo en cuenta sus necesidades desde el principio mediante escenarios o tests. Esto será un punto de partida para nuestra metodología, ya que el usuario de negocio será el centro a partir del cual se construya todo el sistema de análisis y aplicaciones asociadas.

Otro factor a tener en cuenta es el tipo de información con el que nos encontramos. Es muy habitual que la información necesaria se encuentre distribuida entre diferentes fuentes y formatos, y que sea actualizada con su propia frecuencia, además de tener

un volumen que excede las capacidades de las tecnologías tradicionales. Este factor hace que tengamos que ir integrando las diferentes fuentes de forma progresiva, tanto en número de fuentes como en volumen, para ir validando los modelos dentro de un coste sostenido de infraestructura. Además, esta progresión nos ayuda a comprender mejor los factores que afectan a los resultados, y en qué medida. Sería prácticamente inviable tratar de abordar esta clase de problemas complejos directamente, sin hacer una partición previa.

El último factor que suele afectar de forma común a nuestros proyectos es la resistencia al cambio. Es habitual que los equipos de negocio estén acostumbrados a tomar decisiones basándose en pequeños conjuntos de información obtenida de forma manual a través de formularios Excel o informes estáticos. Aunque es posible que un proyecto de este tipo muestre su información en esos formatos, es mucho más potente y conveniente que se habiliten cuadros de mando que los destinatarios de la información puedan manipular para adaptar a sus necesidades en cada momento modificando el periodo de los datos, aplicando filtros, añadiendo comparativas, etc. Es por ello que es necesario formar a los usuarios para que sepan manejar estos cuadros de mando. Esto obliga a que la metodología contemple una fase de formación de usuarios, que no será solo de los cuadros de mando concretos, sino que también incluirá formación sobre Big Data para que puedan empezar a pensar cómo aprovecharlo en sus negocios.

5.2. Resumen de la metodología

Los factores explicados en el apartado 5.1 y el contexto descrito en este tipo de proyectos obligan a que la metodología diseñada tenga una naturaleza iterativa, lo que se traduce en ir entendiendo, cada vez mejor, el contexto en el que se desarrollará el proyecto hasta tener en cuenta todos los factores que influyen en él para luego realizar las iteraciones que sean necesarias para completar el desarrollo, involucrando en todo momento a los *stakeholders* (todas las partes implicadas de la organización o empresa) del proyecto.

Esta metodología toma componentes de la metodología MBDaaS (capítulo 2.1.1) como, por ejemplo, los pares *indicadores/objetivos*, que son los KPI en nuestro diseño, o la estructura básica general de la metodología. También toma como referencia la metodología descrita en el capítulo 2.1.2, ya que se basa en la misma idea de desarrollar el proyecto de forma iterativa para poder captar los requisitos de forma incremental.

No obstante, ninguna de las dos metodologías describe bien cómo adaptarse a un entorno existente, y son demasiado estrictas a la hora de determinar los requisitos. Por este motivo nuestra metodología no solo es iterativa y basada en objetivos, sino que es lo suficientemente flexible y ágil como para permitir que se desarrollen proyectos de Big Data en entornos cuya capa de negocio no tenga la experiencia necesaria como para definir dichos objetivos de forma detallada desde el primer momento, sino que se genere un objetivo general que pueda ser redefinido en futuras iteraciones para contemplar los descubrimientos que se han ido haciendo durante las propias fases de implementación de iteraciones anteriores.

El objetivo es que el proyecto evolucione junto al conocimiento del negocio que se va obteniendo cuando se desarrolla un proyecto de análisis masivo de datos por primera vez. Este es el motivo por el que todo el trabajo se centra en definir de una forma precisa y adecuada los KPI (*Key Performance Indicator* o Indicador Clave de Rendimiento), ya que son los parámetros que establecen, no sólo los requerimientos del

proyecto, sino que también son métricas que podemos utilizar para cuantificar el grado de éxito del proyecto. Es decir, cuando se establece un KPI, la capa de negocio nos está diciendo qué información espera encontrarse en un informe (requisito funcional), y nos está proporcionando un modo de medir si hemos cumplido o no lo que se esperaba del proyecto, que podrá cuantificar si los KPI implementados cubren todos los requisitos y si estos reflejan la información solicitada. Se podría decir que el KPI es el elemento común al lenguaje de los ingenieros Big Data y la capa de negocio.

Este es el resumen de los pasos que componen nuestra metodología:

1. Determinar el objetivo general que se quiere conseguir, quiénes son los implicados y qué información se necesita para acometer el proyecto.

2. Toma de contacto inicial con los interesados en el proyecto.

   a) Captura de requisitos específicos u objetivos que persigue cada interesado.

   b) Documentar el flujo de trabajo que se verá afectado en cada caso.

   c) Documentar el flujo de información dentro del área, especificando las fuentes externas, pero sin entrar en detalle, elaborando el diagrama de flujo de datos, en adelante DFD, de nivel 1.

3. Documentar las fuentes de información del proyecto, estableciendo la relación entre los diferentes flujos de información documentados en un DFD de nivel 2.

4. Diseño de los KPI o métricas clave que se necesitan para cumplir con los objetivos establecidos.

5. Validación y refinamiento de los KPI.

6. Estimación del impacto económico del proyecto.

7. En cada iteración:

   a) Definición de los objetivos de la iteración.

   b) Diseño del modelo de datos que cubra los requisitos especificados.

   c) Obtención de resultados mediante scripting para la validación de los modelos.

   d) Automatización del modelo definido.

   e) Resumen de la iteración y conclusiones.

   f) Formación de los usuarios.

La figura 6 es una representación gráfica de este flujo de trabajo.

5.3. Desarrollo de la metodología

Vamos a estudiar en detalle cada uno de los pasos descritos en el apartado anterior de este trabajo.

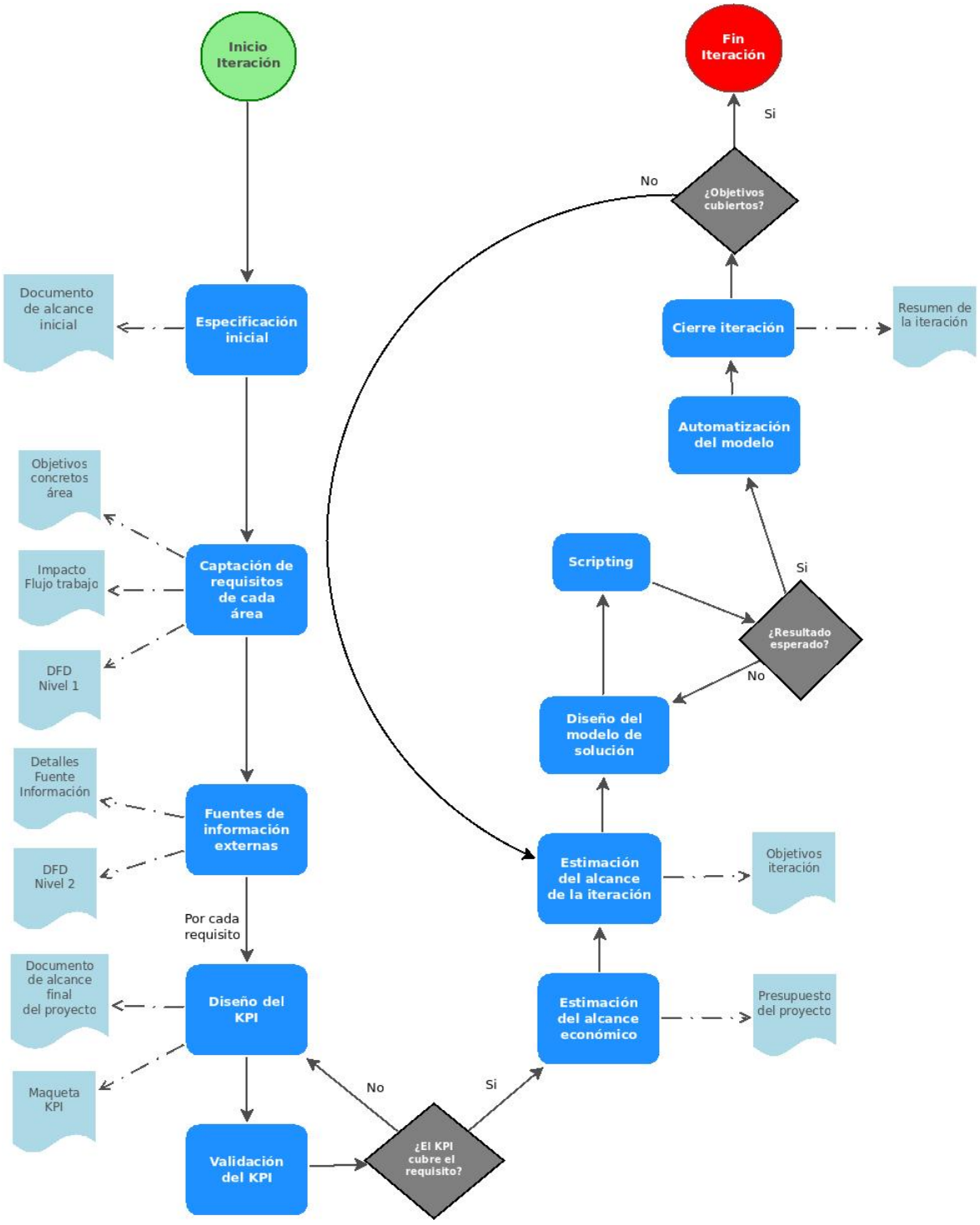


Figura 6: Diagrama de flujo de la metodología

5.3.1. Determinación del objetivo

Comenzaremos el proyecto definiendo los objetivos que se quieren cubrir desde la parte de negocio. En esta fase intervienen los siguientes perfiles profesionales:

- Analista de datos.
- Dirección de la compañía.

En esta primera fase, que puede dividirse en tantas iteraciones como sean necesarias, el analista de datos tiene un papel más pasivo, y recolecta información de los especialistas de negocio que trasladarán la situación actual de la compañía, las áreas que la componen, así como los objetivos de cada área. El objetivo de esta fase es que el analista entienda las necesidades a largo plazo y la estructura de la compañía para comprender la complejidad de los proyectos desarrollados y alinear las soluciones propuestas a estos objetivos. Además, quedarán en evidencia las limitaciones de los sistemas existentes.

Esta fase concluye cuando el analista ha comprendido y documentado qué áreas son estratégicas para la compañía para poder centrarse en conocer los detalles de esta área en fases posteriores.

Al final de esta primera etapa, debe generarse un documento que describa en alcance inicial del proyecto que recoja la siguiente información:

- Áreas o departamentos donde se pretende aplicar el análisis masivo de datos.
- Problemas o retos actuales de dichas áreas, priorizándolos para el valor de la compañía.
- Potencial de las tecnologías Big Data para resolver dichos problemas.
- Estado actual de los datos en cada una de las áreas o departamentos descritos.

5.3.2. Toma de contacto inicial con los interesados

Una vez priorizadas las áreas de actuación del Big Data, tendremos una toma de contacto inicial con los responsables del área. Normalmente, esta implicará al responsable y a un técnico del área.

En esta fase, el papel del analista también es, sobre todo, pasivo ya que debe centrarse en conocer bien el flujo de trabajo del departamento, qué información se utiliza, qué política de datos se sigue, y las limitaciones de la tecnología con la que cuentan actualmente.

Esta fase finaliza cuando el analista ha documentado el flujo de trabajo del área, las fuentes de información utilizadas, y las debilidades que se pretenden cubrir en el futuro proyecto.

Al final de esta área deben generarse los siguientes documentos:

- Documento de objetivos concretos: este documento concreta los problemas descritos en el documento de alcance inicial del proyecto, reflejando las limitaciones técnicas que tiene actualmente la compañía para resolverlos.
- Documento de impacto en los flujos de trabajo: es necesario pensar cómo van a cambiar los procesos para medir el impacto que tendrá el proyecto. Si es posible, se hará una primera propuesta de solución a dicho impacto, proponiendo un nuevo proceso o modificaciones del proceso existente.

- DFD de nivel 1: este diagrama de flujo de datos sólo tiene en cuenta los datos que utiliza y genera el área, sin entrar en detalle de cómo son generados en otras áreas, siendo estos datos externos cajas negras para este nivel. De este modo, entenderemos qué información se necesita, qué se genera, y qué dependencias existen con otros departamentos.

5.3.3. Fuentes de información

En esta fase vamos a profundizar en los detalles de cada una de las fuentes de datos que se utilizan en el área donde se desarrollará el proyecto. De cada fuente de datos conoceremos:

- Origen: la fuente de datos puede ser interna (propia de la compañía) o externa (adquirida a un tercero).
- Modo de uso: cómo se adquiere la información de esta fuente (ApiRest, SOAP, ficheros de texto, etc).
- Frecuencia de consulta.
- Formato.
- Integración con el resto del sistema.

Cuando tenemos claros todos estos puntos, se documentan en un DFD de nivel 2, como mínimo, que muestra el origen y destino de la información que se utiliza en nuestro escenario. Se debe definir un diagrama con los niveles necesarios para comprender los detalles de la información que se está manejando en el proyecto, abriendo las cajas negras que quedaron en el punto anterior hasta el nivel de precisión que sea necesario.

En este punto también se estudia la viabilidad del proyecto desde el punto de vista de los datos, es decir, si contamos con los datos necesarios para cubrir el proyecto o es necesario integrar otras fuentes, como información de otros departamentos o fuentes externas: datos meteorológicos, demográficos, datos de portales abiertos, etc.

5.3.4. Diseño de los KPI

En este punto ya tenemos claros los objetivos que se persiguen en el proyecto y la información que tenemos disponible para cubrirlos. Estos objetivos se definirán con el formato utilizado en la metodología MBDAaaS [6] de "Indicadores/Objetivos", estableciendo el conjunto de KPI's más adecuado para poder cubrirlo.

De cada KPI se debe establecer:

- Objetivo asociado al KPI. Qué objetivo persigue la inclusión de dicho KPI, y cómo apoyará la toma de decisiones en el área de negocio, ya sea mostrando el estado actual de un parámetro importante, avisando cuando se alcanza una situación concreta, etc.
- Qué información lo alimenta y cómo se creará.
- Formato del KPI: puede ser un gráfico, una alerta enviada mediante notificaciones *push*, un dato en un cuadro de mando configurable, etc.

- Disponibilidad del KPI.

Al final de este punto, se debe contar con un documento de alcance final del proyecto, que muestre todas las implicaciones del proyecto que se pretende realizar, así como de la maqueta de cada KPI a desarrollar validada por el negocio y la parte técnica. Esta validación se realizará de forma iterativa para que desde negocio puedan realizar los cambios necesarios para que la aplicación muestre los datos que realmente se necesitan.

5.3.5. Impacto económico del proyecto

Es muy importante establecer el impacto económico del proyecto. Los factores que más influirán en los costes son:

- Adquisición de datos de fuentes de terceros.
- Infraestructura necesaria para acometer el proyecto.
- Contratación de perfiles profesionales específicos, como gestores de infraestructura o ingenieros de datos.
- Impacto en el flujo de trabajo de la empresa.

En esta fase se deben analizar los diferentes proveedores de servicios cloud teniendo en cuenta la experiencia del equipo existente en infraestructuras Big Data. Para estimar el coste real que tiene el proyecto es conveniente realizar pruebas con un subconjunto de datos reales de forma que podamos saber el coste del procesamiento total extrapolando lo que han costado los análisis sobre los conjuntos de prueba.

Todo esto debe servir para elaborar un presupuesto de proyecto, pudiendo ofrecer variantes en función de la tecnología cloud escogida, comparando entre proveedores, opciones de configuración, etc.

5.3.6. Desarrollo de las iteraciones del proyecto

A diferencia de otras metodologías ágiles basadas en iteraciones, en las que estas tienen una duración predeterminada, en nuestra metodología, cada iteración puede tener una duración diferente, en función de la complejidad que se estime, ya que no se establece su duración en base a fechas sino en base a objetivos, ya que son más sencillos de establecer y medir.

Cada iteración comienza concretando los objetivos que se pretenden conseguir, que son traducidos a KPI's. Como se explicó en el apartado 5.3.4, de cada KPI conocemos la información que necesitamos y cómo debe representarse, así que en este punto concretaremos la implementación que se realizará, incluyendo la tecnología, el algoritmo de cálculo, el tiempo de refresco del dato, si la información debe guardarse de modo persistente, qué librería de visualización se utilizará (en caso de ser necesario montar alguna representación gráfica de un dato o KPI), etc.

Teniendo todos estos puntos claros, se realizará un prototipo del KPI mediante prototipado rápido utilizando scripting y visualizaciones básicas para que el negocio pueda ver un primer resultado y validar que es lo que se esperaba. En este punto, estamos a tiempo de hacer cambios en el KPI modificando el script. Una vez validado, se creará un *job* que se ejecute bajo los parámetros establecidos y se creará una visualización definitiva del dato.

Una vez implementados todos los KPI que estaban previstos para esta iteración, se realiza un informe de conclusión indicando qué KPI se han implementado, cómo se están ejecutando, qué informes los utilizan, etc. También se debe indicar si ha habido problemas con algún KPI para valorar si se implementa en siguientes iteraciones, se descarta o se sustituye por uno o varios datos que puedan resultar equivalentes.

Por último, hay que dar una formación a los usuarios sobre interpretación del KPI, asegurándonos que la capa de negocio no tiene problemas para comprender y utilizar el dato en su día a día. Si la capacidad del proyecto lo permite, es posible solapar la fase de formación con la fase de implementación de la siguiente fase del proyecto.

6. Caso práctico: Aplicación de la metodología en Cruz Roja

Antes de poner en práctica nuestra metodología, vamos a presentar el escenario que nos encontramos para dejar claro el contexto en el que la aplicaremos. Entender el entorno en el que se va a trabajar es fundamental para poder desarrollar una solución que encaje con los objetivos de la compañía. Para eso, el especialista debe formar parte del trabajo del día a día, estando completamente integrado con el resto del equipo. Al tratarse de un proyecto académico, no nos es posible estar integrados en el departamento de Cruz Roja, ya que excedería la capacidad de este tipo de trabajos. No obstante, mediante entrevistas y una comunicación activa con el personal de Cruz Roja, se ha modelado y comprendido el funcionamiento de los planes de actuación que se llevan a cabo con las personas que acuden a solicitar asistencia de cualquier tipo. Además, a medida que se avanzaba en el análisis y diseño del proyecto, se validaban los modelos con los especialistas del negocio para asegurarnos de que el trabajo cubre las necesidades reales de nuestro contexto. Esta forma de trabajar nos permite cubrir la carencia de presencia física para poder definir un proyecto que valide el trabajo realizado.

Una vez que hemos establecido los pasos a seguir en nuestra metodología, vamos a implementarla en nuestro escenario real, realizando todas las fases descritas en el apartado 5.2 y generando la documentación requerida en cada una. De este modo, veremos que somos capaces de capturar los requisitos del negocio y desarrollar un proyecto acorde con dichos requisitos.

El escenario escogido para poner a prueba nuestra metodología es Cruz Roja, más concretamente, el área de asistencia a personas en situación de vulnerabilidad. Quizás sea uno de los departamentos donde más información se gestiona y donde más recursos se emplean, por lo que cualquier proyecto que permita optimizar los procesos y que aporte más información a la dirección tendrá un impacto directo en la cuenta de explotación de la compañía.

6.1. Captación de requisitos: Objetivo general

Antes de comenzar con el proyecto, es necesario entender el contexto en el que nos encontramos. Para entenderlo, es necesario describir y entender el estado de madurez de la compañía en cuanto al gobierno del dato, es decir, cómo se gestionan los datos, qué políticas se siguen en su recolección y procesamiento, etc. Todo esto nos ayudará a comprender mejor las necesidades de la compañía y las limitaciones que tendremos a la

hora de aplicar nuestra metodología, previniendo ciertos problemas que puedan surgir durante el desarrollo del proyecto.

Para obtener esta información, nos reunimos con el responsable de TI de la compañía, quien nos indica que cuentan con un gran número de aplicativos, cada uno de los cuales tiene un repositorio propio de datos y, en muchos casos, se conecta a fuentes de datos centrales que almacenan, principalmente, información personal de usuarios. Toda la información se persiste en bases de datos relacionales tradicionales, aunque hay partes que se están migrando a bases de datos no relacionales y aplicativos más modernos.

Como es habitual en organizaciones con sistemas de información antiguos, nos encontramos información fragmentada en diferentes fuentes de información, formatos y frecuencias de refresco. Todo esto complicará la elaboración de nuestro DFD, el cual es muy necesario para comprender el alcance de los datos y cómo se relacionan para poder reorganizar para aprovecharlos mejor.

6.2. Reunión con los *stakeholders*

Comenzamos el proyecto realizando una reunión con el director técnico de Cruz Roja con la intención de concretar en qué área sería más apropiado que se aplicara la metodología propuesta en este trabajo final de máster. El criterio utilizado es que debe ser un área donde exista una gran cantidad de datos, que haya un gran trabajo manual, es decir, que esté poco automatizada, y que, por lo tanto, tenga procesos que puedan ser optimizados gracias al conocimiento generado por este proyecto. El departamento que cumple todos estos requisitos es el *Área de Asistencia a Personas en Situación de Vulnerabilidad*.

Este departamento cumple las condiciones necesarias. Está gobernado por unos procesos claramente definidos, que están bien documentados y son susceptibles de ser analizados para encontrar mejoras una vez que tengamos los datos. También cuenta con una gran cantidad de datos, derivados de los planes que se elaboran para atender a los usuarios, los PPI (Ver apartado 6.3), que registran a la perfección la situación actual de un usuario y los planes que se llevarán a cabo para mejorarla. Lo interesante es su naturaleza, que es el motivo por el que lo elegimos.

Se trata de un departamento que trabaja directamente con las personas con problemas, tratando de ayudarles a superar su situación actual. Los datos que se generan son derivados de conversaciones y relaciones personales, por lo que son mucho más subjetivos que los que encontramos en negocios de otros ámbitos, como puede ser la banca, donde se registran hechos objetivos (transacciones bancarias, estado de cuentas, etc.). En este caso, nos encontramos con usuarios que pueden dar datos ambiguos, o que mienten, o que tienen situaciones que cambian con el tiempo. La propia naturaleza de este trabajo implica que se haga de forma manual. Se establecen relaciones personales estrechas con los usuarios con el fin de aumentar las probabilidades de éxito de los proyectos. Los trabajadores sociales están acostumbrados a basarse en su amplia experiencia para tomar decisiones, pero no tienen ninguna herramienta que los asista, de forma objetiva, en esta toma de decisiones.

Todos estos factores propician que el director técnico proponga el área de intervención para desarrollar el proyecto.

6.3. Marco de Atención a las Personas (MAP)

Para llevar a cabo el proyecto, es necesario comprender la metodología de trabajo de Cruz Roja. Para ello, se ha concertado una reunión con la persona que da las formaciones internas del PPI (Plan Personalizado de Interfención) en Cruz Roja para que nos explique todos los detalles de este proceso tan importante.

La metodología que define el modo de trabajar de este área es el Marco de Atención a las Personas, o MAP por sus siglas. Esta metodología define los procesos, los materiales empleados, los perfiles y las competencias necesarias para llevar a cabo todo el trabajo del departamento. Esta metodología está descrita en la figura 7, que muestra las fases de acogida, valoración, ejecución, donde entra el PPI, la evaluación de la satisfacción y el compromiso con Cruz Roja.

Cuando un usuario acude a Cruz Roja Alicante, en adelante CRA, en busca de ayuda porque se encuentra en una situación de vulnerabilidad, un agente social elabora un completo informe para realizar una foto de la situación actual de dicha persona. A partir de este momento, este establece todas las acciones que van a llevarse a cabo para ayudar al usuario desde cualquier área de CRA. Estas actuaciones pueden ser desde seguimientos telefónicos o acompañamiento con voluntarios hasta planes integrales de formación o aportaciones económicas de distinto grado. Es importante destacar la gran variedad de actuaciones que pueden realizarse ya que hacen que el sistema que registra toda la información sea ciertamente complejo.

Una vez establecido el plan, se lleva a cabo la ejecución del mismo, donde se realizan las acciones programadas. Como hemos dicho, hay una gran variedad de acciones que se realizan en la asistencia a usuarios. Es posible que algunos usuarios reciban una cierta cantidad económica, que otros usuarios reciban una llamada para recordarles que se tomen la medicación e interesarse por su estado, otro usuario puede recibir la visita de un voluntario y que lo acompañe al médico, y así podríamos seguir con muchos más ejemplos. Además, estas actuaciones pueden ser puntuales o periódicas, e involucrar a varias áreas de Cruz Roja.

Cuando el trabajador de CRA lo considera oportuno, puede cerrar el caso y hacer un nuevo informe completo para tomar una nueva foto de la situación actual del usuario con el fin de establecer si las actuaciones realizadas han tenido el impacto esperado y decidir si es necesario realizar nuevas acciones o se considera que la situación de vulnerabilidad que empujó al usuario a solicitar ayuda ha sido solventada. Si es necesario continuar con el proyecto, dicho informe servirá para definir nuevas acciones y comenzar de nuevo. Si no, se cerrará el expediente, con lo que este informe servirá de resumen de la mejora de la situación del usuario.

Todo este proceso está descrito en la figura 7, y ha supuesto una revolución en Cruz Roja, que ha pasado de centrarse en los proyectos a centrarse en los usuarios. El enfoque por proyectos consiste en elaborar planes de actuación sueltos que cubran necesidades concretas de un usuario. En cambio, en el nuevo enfoque, basado en el usuario, se trata de analizar la situación global de la persona para, a partir de ahí, coordinar las diferentes acciones para optimizar recursos y mejorar los resultados.

Además, el enfoque por usuarios, desde el punto de vista de los datos, nos resulta muy interesante, porque nos permite analizar la situación de todos los individuos para, dado un nuevo caso, poder clasificarlo para recomendar acciones que hayan sido determinantes en el éxito de los proyectos de asistencia.

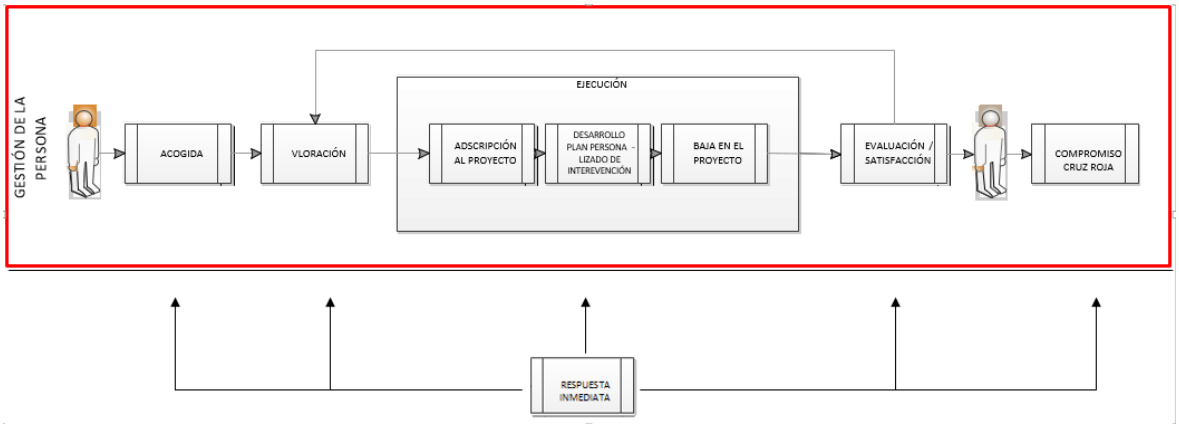


Figura 7: Diagrama de la metodología MAP de Cruz Roja (Fuente: Cruz Roja)

6.4. Aplicación de la metodología

Una vez que comprendemos nuestro contexto, vamos a aplicar la metodología en dicho contexto para validarla. Dado que se trata de un trabajo académico, y por motivos de tiempo, solo vamos a desarrollar una iteración. Esto es suficiente para ver que la metodología es capaz de capturar las necesidades del negocio de la forma adecuada y diseñar un proyecto acorde a dichas necesidades, sin entrar en toda la fase de implementación, que quedaría fuera del ámbito de este trabajo.

6.4.1. Fuentes de información del proyecto

Cruz Roja Alicante cuenta con multitud de fuentes de información. Algunas de ellas son aisladas, pero la mayoría trabajan entre sí para cubrir los diferentes servicios. Hay multitud de servicios que se han desarrollado de forma independiente, aprovechando algo de información existente pero, sobre todo, y en la mayoría de casos, creando su propio repositorio de datos para poder operar. Esto ha provocado que exista un ecosistema de datos muy complejo, cuyo análisis completo mediante un DFD llevaría demasiado tiempo como para ser abordado por completo en este proyecto. De este modo, desde el área de TI de la organización, han generado un repositorio de datos agrupando información del área de acogida para que podamos hacer un análisis preliminar de dichos datos. Esta información, en el contexto original, está distribuida entre varias bases de datos.

En un proyecto real, habría que definir un diagrama de flujo de datos para documentar y comprender de qué información se dispone, dónde está almacenada y cómo fluye entre los diferentes aplicativos o departamentos. Dado que en este proyecto contamos con un único repositorio de datos, vamos a describirlo pero no vamos a generar un DFD porque no conocemos los detalles del entorno real de Cruz Roja.

Lo que pretendemos con este análisis es conocer la calidad del dato, su completitud y establecer mejoras que pueden hacerse del repositorio en sucesivas iteraciones, además de comprender qué tipo de información se guarda asociada a los usuarios del área de acogida.

El dataset de prueba que nos ha facilitado Cruz Roja está subdividido en tres tablas correspondientes a tres consultas a su base de datos.

Dado que no conocemos la estructura de información real de Cruz roja, no podemos elaborar un DFD, pero sí que podemos crear unas tablas que muestren los detalles de

los datos y que nos ayuden a entender mejor el tipo de información que podemos extraer de ellos. Para ello, hemos elaborado las tablas 1, 2 y 3, que se pueden encontrar en el anexo A, y que muestran las características básicas de los datos de las tres consultas que nos han facilitado desde la organización.

La primera consulta, descrita en la tabla 1, contiene la información de las valoraciones recogidas por los técnicos, de modo que son una foto de la situación de los usuarios en el momento de la valoración, que queda descrita por multitud de campos que tratan de capturar los detalles importantes de cualquier aspecto de la vida de los usuarios. La segunda consulta, cuyos campos están descritos en la tabla 2, recoge los planes desarrollados con los usuarios, y están relacionados con la situación descrita en la tabla 1 mediante un identificador de usuario. Estos planes son elaborados a partir de la información recogida en las valoraciones. Por último, la tabla 3 recoge las actividades que se han llevado a cabo para ejecutar los planes de la consulta 2.

Este dataset contiene algunos campos que no están informados y otros campos con un campo de texto libre, que no será analizado en esta primera iteración ya que requiere de técnicas avanzadas de procesamiento de lenguaje natural. Por ello, es necesario realizar un preprocesamiento de los datos para que los algoritmos funcionen correctamente. Este preprocesamiento consiste en el descarte de los campos que tengan menos de un 50 % de completitud así como los que sean de tipo *string*, ya que contienen texto libre. Además, antes de aplicar cualquier algoritmo sobre estos datos aplicaremos algoritmos de reducción de dimensionalidad, como PCA, para encontrar columnas que sean redundantes. El preprocesamiento de los datos persigue dos objetivos fundamentales, que los cálculos sera mucho más rápidos y que, al mismo tiempo, podamos valorar dejar de recoger datos que no aportan nueva información. Este análisis es extrapolable a las otras dos consultas, por lo que se descartan directamente las columnas no informadas.

6.4.2. KPI's: Diseño, refinamiento y validación

Como hemos visto en el capítulo 6.3, el área de acogida de Cruz Roja trabaja con mucha información, de la cual se extraen estadísticas básicas. En las entrevistas realizadas durante la captación de requisitos se indicó que no se está realizando ningún análisis profundo que asista a los trabajadores sociales a la hora de tomar decisiones ni a la capa de negocio a comprender mejor los recursos que se demandan en este área. En el caso de los trabajadores sociales, estos se guían por informes solicitados de forma puntual de ciertos datos de los PPI y, en el caso de la capa de negocio es algo parecido, pero con un conjunto de datos diferente, por lo que se han establecido una serie de KPI que pretenden ser la base de un conocimiento más profundo mediante KPI más complejos.

- KPI's para la capa de negocio:
  - Número de perfiles de usuario.
  - Número de usuarios en cada uno de los perfiles.
  - Número de proyectos realizados en cada perfil y en total.
- KPI's para los trabajadores sociales:
  - Características propias de cada perfil.

Vamos a ver todos los KPI más en detalle:

Número de perfiles de usuario. Mediante un algoritmo de clasificación no supervisado como son los mapas auto organizados (o *SOM*, por sus siglas en inglés), vamos a clasficar los datos de nuestro conjunto para agrupar usuarios similares (Ver capítulo 6.4.6). Este KPI mostrará cuántos tipos de usuario acuden al servicio.

Detalles del KPI:

- Objetivo: entender cuántos tipos de usuarios diferentes acuden al servicio de asistencia de Cruz Roja.
- Datos: este KPI se basa en los datos de los perfiles obtenidos mediante el Marco de Atención a las Personas (MAP, capítulo 6.3), que reflejan con la máxima precisión la situación actual de los usuarios.
- Formato: documento elaborado de forma periódica (proceso batch).
- Disponibilidad: el KPI se calculará de forma mensual.

Número de usuarios en cada uno de los perfiles. Una vez que tengamos los perfiles, estableceremos cuántos usuarios hay en cada uno de ellos. Si analizamos este dato con el tiempo, podemos ver cómo evoluciona, si hay estacionalidad, etc.

Detalles del KPI:

- Objetivo: estudiar el volumen de cada uno de los perfiles puede utilizarse, en sucesivas fases del proyecto, para hacer previsiones de materiales, personal, etc.
- Datos: está basado en los mismos datos que el KPI "Número de perfiles de usuario".
- Formato: documento elaborado de forma periódica (proceso batch).
- Disponibilidad: el documento se generará mensualmente.

Número de proyectos realizados en cada perfil y en total. Es importante para la capa de negocio conocer cuántos proyectos se llevan a cabo para cada perfil con el objetivo de conocer mejor cómo se invierten los recursos, dónde se puede mejorar, etc.

Detalles del KPI:

- Objetivo: estimar el esfuerzo realizado en cada uno de los grupos mediante el número de proyectos realizados en cada grupo.
- Datos: se basa en los mismos datos que el KPI "Número de perfiles de usuario".
- Formato: documento elaborado de forma periódica (proceso batch).
- Disponibilidad: el documento se generará mensualmente.

Características propias de cada perfil. Lo que más le interesa a un trabajador social es entender las similitudes que existen entre los diferentes tipos de usuario, y para eso lo mejor es que vea qué perfiles hay y qué características son propias de cada uno. Para entender bien las características de cada grupo, se pude añadir mucha información a cada usuario, tanto de otras fuentes internas de Cruz Roja, que provendrían de bases de datos de otras aplicaciones, como de fuentes externas como datos abiertos del INE, índices demográficos, etc. No obstante, el estudio que se puede hacer con los datos actuales es muy interesante, y puede sentar una buena base para una primera comprensión de nuestros usuarios.

Detalles del KPI:

- Objetivo: conocer las características propias de los usuarios pertenecientes a cada uno de los grupos detectados en el KPI "Número de perfiles de usuario".
- Datos: se basa en los mismos datos que el KPI "Número de perfiles de usuario".
- Formato: documento elaborado de forma periódica (proceso batch).
- Disponibilidad: el documento se generará mensualmente.

Este grupo de KPI no tiene como finalidad aportar un conocimiento profundo sobre el negocio, ya que están basados en un conjunto de datos muy reducido. Su valor reside en que son la base para comenzar a arrojar algo de luz sobre el potencial de los datos que tiene el área de atención a usuarios, ayudando a definir las fases posteriores de este proyecto, que tratarán de calcular KPI más complejos y útiles para la capa de negocio. Por este motivo tampoco se ha definido una visualización específica para ellos, porque los datos realmente interesantes para el negocio son esos datos más complejos basados en este primer conjunto.

Esta aproximación se corresponde con la fase de estimación de objetivos de la iteración de nuestra metodología, en la que establecemos la información que queremos obtener de los datos a partir de la información existente (KPI diseñados en iteraciones anteriores) y los nuevos datos que se pueden aportar en la nueva iteración.

6.4.3. Impacto económico de la iteración

Esta iteración tiene como objetivo el descubrimiento y la preparación de iteraciones posteriores, por lo que se suelen trabajar con conjuntos pequeños de datos y con sistemas de scripting, más que con sistemas en producción y grandes datasets. Por este motivo, el coste de infraestructura a estas alturas de proyecto es muy bajo. Dado que estamos trabajando en un entorno *cloud* y que trabajamos con un dataset que no llega a los 300MB, vamos a hacer una reserva de 200 € para ejecutar los procesamientos de scripts que sirven para la exploración de datos y la validación de los modelos iniciales.

En esta iteración inicial no se adquirirán datos de fuentes externas que supongan un coste adicional, debido a que se está haciendo el descubrimiento inicial y el planteamiento de los primeros KPI. Tampoco se valora un impacto en los procesos existentes en la empresa que supongan un coste económico, ya que estos cambios son el resultado de análisis más complejos realizados en iteraciones avanzadas del proyecto.

Sí que es necesario contar con un especialista en infraestructuras cloud, un ingeniero de datos y un analista de datos, por lo que vamos a asignar un presupuesto de 10.000 € a esta primera iteración para el personal encargado del análisis y la implementación.

6.4.4. Objetivos de la iteración

Los objetivos de esta primera iteración son:

- Conocer los perfiles de usuario que acuden a solicitar ayuda al área de atención.
- Establecer un primer informe que pueda mostrar todos los KPI descritos en el capítulo anterior.
- Realizar un primer análisis de los datos proporcionados por Cruz Roja.

6.4.5. Datos necesarios para la iteración

Los únicos datos con los que contamos para esta iteración es el fichero descrito en el capitulo 6.4.1 de este documento. En iteraciones posteriores del proyecto, como se describe en las conclusiones del mismo, se puede incorporar nueva información que añada más posibilidades de análisis, permitiendo diseñar nuevos KPI.

6.4.6. Diseño y validación de los modelos de datos

En esta primera iteración, sólo necesitamos elaborar un modelo de datos capaz de realizar la clasificación de nuestro conjunto de prueba, ya que los KPI definidos no nos exigen un modelo más complejo. No obstante, cabe esperar que en futuras iteraciones, donde se añadan más fuentes de datos y se definan nuevos KPI, haya que ampliar este modelo inicial, bien con la creación de nuevos modelos o con su refactorización, para dar cabida a los nuevos requerimientos.

El modelo de inteligencia artificial escogido para realizar la clasificación no supervisada de nuestros usuarios en base a la información disponible por sus entrevistas es el algoritmo de SOM, que permite cumplir los KPI de agrupación y caracterización de los grupos de usuarios de la organización en conjuntos de datos no etiquetados, como es el nuestro [1].

El algoritmo SOM es muy potente y efectivo para resolver estos problemas. Se basa en redes neuronales multicapa capaces de definir fronteras entre los elementos del conjunto de datos basándose en la información agregada de todas sus columnas.

No obstante, primero se preprocesarán los datos aplicando PCA para eliminar columnas redundantes, obteniendo el modelo de datos representado en la figura 8.

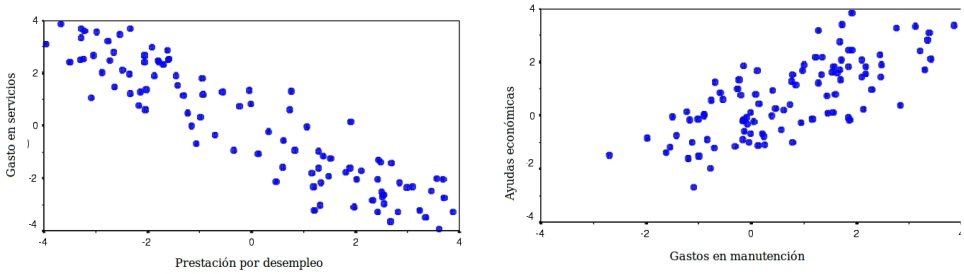
Figura 8: Modelo de datos de la primera fase del proyecto

Después de haber realizado el PCA, se ha calculado la matriz de correlación para los parámetros del dataset, obteniendo resultados mostrados en las gráficas de la figura 9. Es importante tener en cuenta que las gráficas que pueden verse en esta imagen sólo representan un subconjunto de los datos totales con los que cuenta Cruz Roja normalizados para ajustar las escalas. Esto es debido a que se agrupan los datos cercanos entre sí a la hora de representarlos para verlos más claramente, y a que no todas las columnas tienen todos los datos. Con esta agrupación, conseguimos una visualización

mejor de los datos que nos permite identificar, de forma visual, si dos columnas están o no relacionadas.

La gráfica de la figura 9a muestra, en el eje Y los gastos que los usuarios dedican a servicios en el hogar, mientras que el eje X muestra la prestación por desempleo recibida. El hecho de que exista una correlación lineal negativa, como muestra la figura, entre estas dos variables indica que a medida que que la prestación en desempleo crece, descienden los gastos en servicios de los usuarios. Sería interesante, en fases posteriores del proyecto, analizar el resto de variables para ver por qué existe esta relación y si puede elaborarse un KPI interesante basado en esta información.

En la figura 9b podemos ver una correlación lineal positiva entre las ayudas económicas que se fijan en los PPI, es decir, el dinero que se da a los usuarios como parte de un PPI, y los gastos en manutención de los usuarios. Esto indica que cuanto más gasto en manutención tienen los usuarios, más dinero reciben, lo que además indica que gran parte de este dinero va destinado a cubrir los gastos de manutención de las familias.



(a) (b)

Figura 9: Correlaciones entre variables del dataset

La figura 10 muestra el resultado obtenido con el SOM. Como puede observarse, el algoritmo no ha sido capaz de clasificar a los usuarios de forma clara. En lugar de ello, los ha repartido de forma más o menos equitativa entre todos los grupos de la matriz de 20x20. Por lo tanto, no es posible establecer una clasificación entre los usuarios con los datos de prueba, por lo que es probable que sea necesario añadir nuevos datos que permitan que la red neuronal pueda segmentar a los usuarios.

Dado que no podemos realizar una clasificación con los datos actuales, no podremos dar unos valores concretos para los KPI propuestos en esta iteración. No obstante, procesaremos los datos para orientar la importación de nuevos datos en iteraciones posteriores.

6.4.7. Automatización del modelo

Para automatizar el cálculo, no solo del modelo de SOM propuesto en esta iteración, sino los futuros modelos, tanto en tiempo real como en *batch* que se propongan en futuras iteraciones, se creará un sistema con una arquitectura Lambda (figura 11). Esta arquitectura está pensada para combinar fuentes de datos estáticas como bases de datos SQL, ficheros de datos en formatos como CSV, JSON o XML, entre otros, o logs de diferentes servidores [21]. Como puede verse en el esquema, típicamente las fuentes de datos estáticas se conectarán directamente al almacenamiento HDFS, de mayor tamaño, para procesamientos posteriores. En cambio, los logs, además de guardarse en esta capa de HDFS (en cuyo caso pueden guardarse en bruto o con algún tipo de

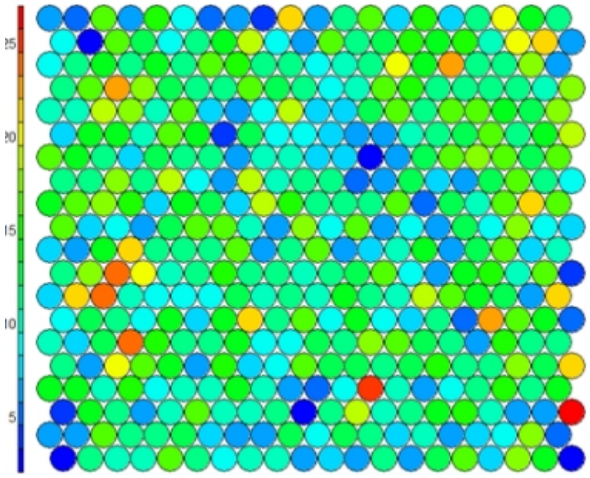


Figura 10: Matriz de resultados del SOM

preprocesamiento previo que reduzca su tamaño), serán volcados en un servidor Kafka para alimentar las vistas incrementales utilizando procesamientos en tiempo real.

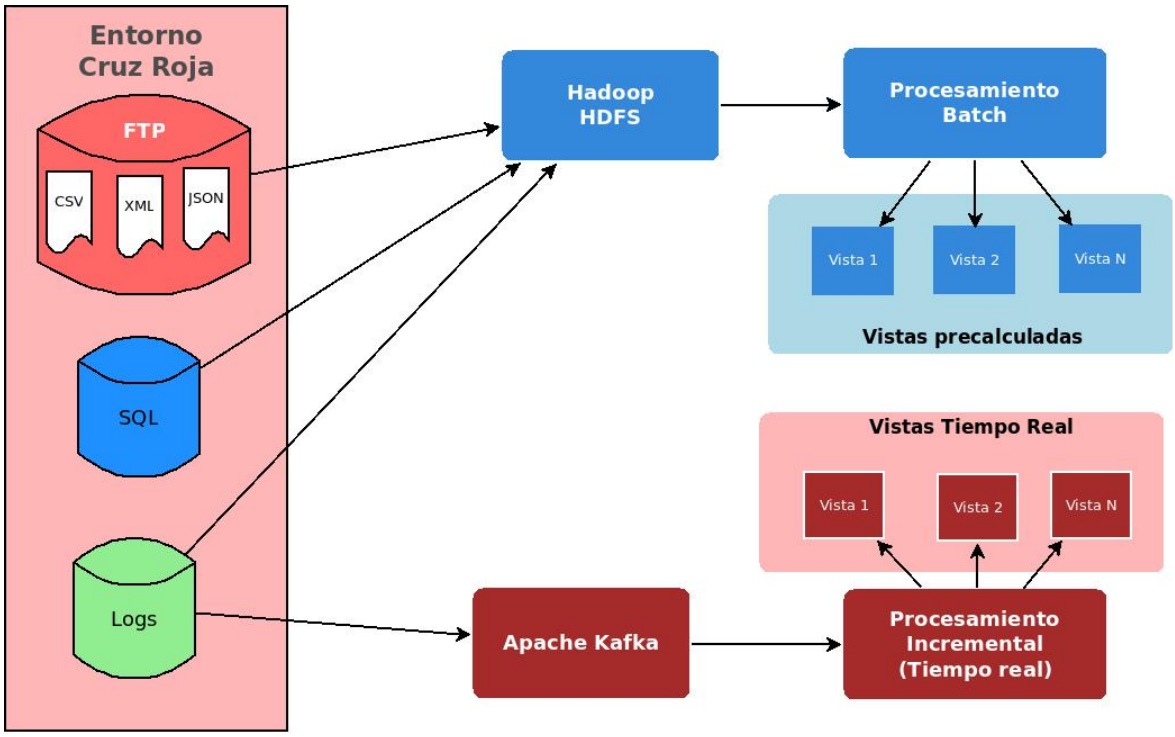


Figura 11: Arquitectura Lambda de ingesta y procesamiento de datos en Cruz Roja

De este modo, podemos tener dos tipos de visualizaciones: Por un lado, podemos tener visualizaciones precalculadas, que muestren datos del pasado, quizás de periodos relativamente largos de tiempo, que sirvan para observar patrones estacionales o resultados de estrategias durante periodos largos de tiempo y, por otro lado, podemos tener visualizaciones más cercanas al tiempo real que muestren la situación actual de la compañía, especialmente útil para mostrar indicadores críticos a corto plazo que ayuden a tomar decisiones muy rápidas.

Otra de las ventajas clave que nos hacen inclinarnos por esta arquitectura es que se pueden integrar nuevas fuentes de datos de forma progresiva. Para ello, simplemente

será necesario conectarlas a Kafka o a HDFS (existen conectores para los tipos de servidores más habituales, o puede desarrollarse uno propio), crear los procesamientos asociados a los nuevos datos y las visualizaciones donde se mostrará el resultado.

En cuanto a la parte de visualización, proponemos en la figura 12 para mostrar la información de cada grupo de usuarios. Este dashboard solo mostrará datos de nuestro dataset de pruebas, de modo que no es necesario actualizarlo periódicamente, sino que será generado por la rama *Batch* de la arquitectura lambda para crear las vistas. Este dashboard mostrará datos de 3 de los 4 KPI propuestos, ya que el KPI "Características propias de cada perfil" debe ser analizado mediante la observación de los parámetros de la red neuronal, ya que no es posible una visualización directa del resultado de la clasificación realizada con el SOM (Capítulo 6.4.6), sino que hay que observar qué variables son representativas de cada uno de los conjuntos.

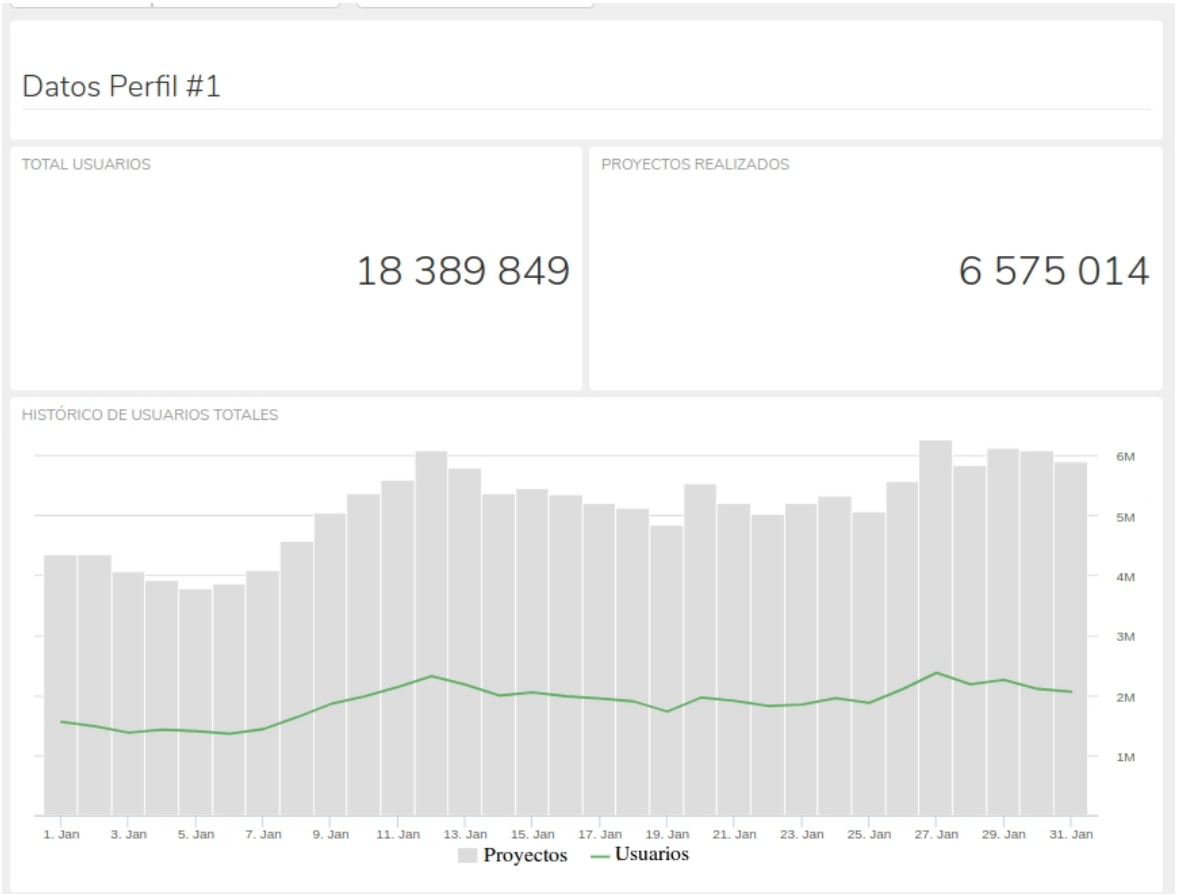


Figura 12: Plantilla del dashboard para mostrar los KPI de esta iteración (Datos ficticios)

6.4.8. Resumen de la iteración

En esta iteración hemos conseguido definir cuál será nuestra arquitectura de procesamiento de datos, hemos tomado contacto con los datos procesando el dataset de pruebas proporcionado por Cruz Roja, y hemos conseguido definir y procesar unos primeros KPI.

Para iteraciones posteriores, es necesario definir KPI más ambiciosos que complementen a los definidos hasta ahora con información de nuevas fuentes de datos como,

por ejemplo, datos de coste económico de los proyectos, datos demográficos, etc. así como otras fuentes de datos internas de la compañía. Es importante elaborar, como indica la metodología, un DFD que muestre las relaciones entre las diferentes fuentes de datos que se van incorporando.

También sería conveniente prescindir del conjunto de datos inicial de prueba y conectar a la arquitectura lambda definida las fuentes de datos reales para poder empezar a trabajar con información más cercana a la realidad.

Por último, en cuanto a los datos, se pueden aplicar técnicas de procesamiento de lenguaje natural a los campos que sean de tipo *string* para extraer información interesante de ellos y complementar a los datos ya procesados.

6.4.9. Formación de los usuarios

Se ha establecido la necesidad de impartir una formación específica a los usuarios sobre las tecnologías más utilizadas en proyectos Big Data, como las mencionadas en el marco teórico de este proyecto, sección 2, para que entiendan el potencial de este nuevo paradigma. También se explicarán casos de uso en contextos similares, así como los objetivos que se pretendían alcanzar en esta iteración.

Con todo esto, se pretende que los usuarios tengan un mayor conocimiento del Big Data para que estén más implicados en el proyecto y sean capaces de poner todo su conocimiento de negocio al servicio de la tecnología y el proyecto.

En futuras iteraciones, esta fase servirá para que los usuarios conozcan las nuevas tecnologías que se han introducido en esa iteración, los nuevos KPI y objetivos cubiertos, y el impacto generado en la compañía. Además, si es necesario, se les enseñará a manejar los nuevos cuadros de mando implementados utilizando ejemplos de aplicaciones reales para que aprovechen todo su potencial.

La fase de formación de los usuarios es muy importante, ya que los usuarios se enfrentan a cuadros de mando y conceptos muy complejos, y es necesario que entiendan cómo estas tecnologías y técnicas están alineadas con su trabajo del día a día, es decir, con las estrategias de negocio.

7. Conclusiones

En este proyecto hemos conseguido formalizar una metodología capaz de implantar un proyecto de Big Data en organizaciones pequeñas o con poca experiencia en este tipo de proyectos. Además, hemos conseguido llevar a cabo una prueba piloto en un entorno real que nos ha permitido refinar y validar el desarrollo obtenido. Ha quedado patente la mejora conseguida respecto a las metodologías existentes descritas en el marco teórico, y esperamos haber sentado una base para el desarrollo de un verdadero estándar que sirva como referencia para proyectos de este tipo, cada vez más comunes en empresas y organizaciones de todo tipo. Por lo tanto, damos por cumplido nuestro objetivo general.

En cuanto a los objetivos específicos, hemos cubierto todos ellos ya que hemos estudiado las metodologías Big Data más relevantes, diseñadas para diferentes propósitos, lo que nos ha permitido elaborar la nuestra propia aprovechando las fortalezas de cada una de ellas, y la hemos validado aplicándola en el entorno real gracias a Cruz Roja. Todo esto nos ha servido para reflexionar sobre metodologías ágiles, modelos iterativos, guiar las decisiones de las compañías por los datos, y otros aspectos de este tipo de proyectos.

En cuanto a la metodología propuesta, como futuras mejoras o desarrollos, sería conveniente buscar estándares adecuados para el documento de alcance inicial y el de impacto económico del proyecto que, por motivos de tiempo, no han sido especificados en la metodología, aunque sí que se han indicado como necesarios, elaborando una versión libre en la aplicación con Cruz Roja.

Como trabajo futuro, sería necesario finalizar la implantación en Cruz Roja llevando a cabo todas las fases necesarias para conseguir un producto funcional, al menos, en el área de atención a personas vulnerables. De este modo, se podrán refactorizar los detalles de la metodología propuesta en este trabajo. Además, es necesario implementar la metodología en otros entornos para refinarla buscando la máxima generalidad, y eliminando los posibles acoplamientos que se hayan podido generar debido al entorno en el que se ha modelado. Además, si se termina estableciendo un estándar, ya sea en el objetivo general de implantar un sistema de Big Data en empresas o en algunos de los pasos que componen cada iteración, será necesario modificar la metodología para adaptarla a dichos estándares.

Bibliografía y referencias

[1] U. F. Alias, N. B. Ahmad y S. Hasan. «Mining of E-learning behavior using SOM clustering». En: *2017 6th ICT International Student Project Conference (ICT-ISPC)*. 2017 6th ICT International Student Project Conference (ICT-ISPC). Mayo de 2017, págs. 1-4. DOI: 10.1109/ICT-ISPC.2017.8075350.

[2] Hamza Hussein Altarturi, Keng-Yap Ng y Mohd Izuan Hafez Ninggal. «A requirement engineering model for big data software». En: *2017 IEEE Conference on Big Data and Analytics (ICBDA)* (2017). DOI: https://doi.org/10.1109/ICBDAA.2017.8284116.

[3] *Apache Hadoop Website*. URL: http://hadoop.apache.org.

[4] *Apache Parquet*. URL: https://parquet.apache.org/ (visitado 29-01-2019).

[5] *Apache Spark Website*. URL: http://spark.apache.org.

[6] C. A. Ardagna y col. «A Model-Driven Methodology for Big Data Analytics-as-a-Service». En: *2017 IEEE International Congress on Big Data (BigData Congress)*. 2017 IEEE International Congress on Big Data (BigData Congress). Jun. de 2017, págs. 105-112. DOI: 10.1109/BigDataCongress.2017.23.

[7] D. Becker, T. D. King y B. McMullen. «Big data, big data quality problem». En: *2015 IEEE International Conference on Big Data (Big Data)*. 2015 IEEE International Conference on Big Data (Big Data). Oct. de 2015, págs. 2644-2653. DOI: 10.1109/BigData.2015.7364064.

[8] Thomas C. Bressoud y Qiuyi Tang. «Results of a Model for Hadoop YARN MapReduce Tasks». En: *2016 IEEE International Conference on Cluster Computing (CLUSTER)* (2016). DOI: https://doi.org/10.1109/CLUSTER.2016.51.

[9] T. Chin y D. Suter. «Incremental Kernel Principal Component Analysis». En: *IEEE Transactions on Image Processing* 16.6 (jun. de 2007), págs. 1662-1674. ISSN: 1057-7149. DOI: 10.1109/TIP.2007.896668.

[10] © 2018 Cloudera y col. *Machine Learning — Analytics — Cloud*. Cloudera. URL: https://www.cloudera.com/ (visitado 25-12-2018).

[11] *Data Management Platform, Solutions and Big Data Analysis*. Hortonworks. URL: https://es.hortonworks.com/ (visitado 25-12-2018).

[12] Tom DeMarco. *Structured Analysis and System Specification*. Yourdon Press, 1978. ISBN: 0-917072-07-3.

[13] P. Gonzalez-Alonso, R. Vilar y F. Lupiañez-Villanueva. «Meeting Technology and Methodology into Health Big Data Analytics Scenarios». En: *2017 IEEE 30th International Symposium on Computer-Based Medical Systems* (2017). DOI: https://doi.org/10.1109/CBMS.2017.71.

[14] Georgios Gousios. «Big Data Software Analytics with Apache Spark». En: *2018 IEEE/ACM 40th International Conference on Software Engineering: Companion (ICSE-Companion)* (2018). DOI: https://doi.org/10.1145/3183440.3183458.

[15] R. He y col. «Robust Principal Component Analysis Based on Maximum Correntropy Criterion». En: *IEEE Transactions on Image Processing* 20.6 (jun. de 2011), págs. 1485-1494. ISSN: 1057-7149. DOI: 10.1109/TIP.2010.2103949.

[16] A. Imawan y J. Kwon. «A timeline visualization system for road traffic big data». En: *2015 IEEE International Conference on Big Data (Big Data)*. 2015 IEEE International Conference on Big Data (Big Data). Oct. de 2015, págs. 2928-2929. DOI: 10.1109/BigData.2015.7364125.

[17] «ISO/IEC/IEEE Approved Draft International Standard - Systems and Software Engineering – Life Cycle Processes –Requirements Engineering». En: *ISO/IEC/IEEE P29148_FDIS, September 2018* (ene. de 2018), págs. 1-104.

[18] Atif Aftab Ahmed Jilani, Aamer Nadeem y Tai-hoon Kim. «Formal Representations of the Data Flow Diagram: A Survey». En: *2008 Advanced Software Engineering and Its Applications* (2008). DOI: https://doi.org/10.1109/ASEA.2008.34.

[19] Romeo Kienzler. *Mastering Apache Spark 2.x*. Packt Publishing, 2017. ISBN: 978-1-78646-274-9.

[20] U. A. Kumar e Y. Dhamija. «Comparative analysis of SOM neural network with K-means clustering algorithm». En: *2010 IEEE International Conference on Management of Innovation Technology*. 2010 IEEE International Conference on Management of Innovation Technology. Jun. de 2010, págs. 55-59. DOI: 10.1109/ICMIT.2010.5492838.

[21] Nathan Marz y James Warren. *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications, abr. de 2015. 328 págs. ISBN: 978-1-61729-034-3. URL: https://www.manning.com/books/big-data.

[22] J. McHugh y col. «Integrated access to big data polystores through a knowledge-driven framework». En: *2017 IEEE International Conference on Big Data (Big Data)*. 2017 IEEE International Conference on Big Data (Big Data). Dic. de 2017, págs. 1494-1503. DOI: 10.1109/BigData.2017.8258083.

[23] A. Mjeda y M. Hinchey. «Requirement-centric Reactive Testing for Safety-Related Automotive Software». En: *2015 IEEE/ACM 2nd International Workshop on Requirements Engineering and Testing*. 2015 IEEE/ACM 2nd International Workshop on Requirements Engineering and Testing. Mayo de 2015, págs. 5-8. DOI: 10.1109/RET.2015.9.

[24] D. Mougouei. «Factoring requirement dependencies in software requirement selection using graphs and integer programming». En: *2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE)*. 2016 31st IEEE/ACM International Conference on Automated Software Engineering (ASE). Sep. de 2016, págs. 884-887.

[25] Roger S Pressman. *Software engineering: a practitioner's approach*. 6th ed., International ed. Boston [etc.]: McGraw-Hill Higher Education, 2005. 912 págs. ISBN: 978-0-07-123840-3.

[26] Jeffrey S. Saltz. «The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness». En: *2015 IEEE International Conference on Big Data* (2015). DOI: https://doi.org/10.1109/BigData.2015.7363988.

[27] Jeffrey S. Saltz e Ivan Shamshurin. «Big data team process methodologies: A literature review and the identification of key factors for a project's success». En: *2016 IEEE International Conference on Big Data (Big Data)* (2016). DOI: https://doi.org/10.1109/BigData.2016.7840936.

[28] Garima Sharma y Anita Ganpati. «Performance evaluation of fair and capacity scheduling in Hadoop YARN». En: *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)* (2015). DOI: https://doi.org/10.1109/ICGCIoT.2015.7380591.

[29] S. Sivanandan y Yogeesha C. B. «Agile development cycle: Approach to design an effective Model Based Testing with Behaviour driven automation framework». En: *20th Annual International Conference on Advanced Computing and Communications (ADCOM)*. 20th Annual International Conference on Advanced Computing and Communications (ADCOM). Sep. de 2014, págs. 22-25. DOI: 10.1109/ADCOM.2014.7103243.

[30] H. K. Solvang y col. «Frequency-Domain Pearson Distribution Approach for Independent Component Analysis (FD-Pearson-ICA) in Blind Source Separation». En: *IEEE Transactions on Audio, Speech, and Language Processing* 17.4 (mayo de 2009), págs. 639-649. ISSN: 1558-7916. DOI: 10.1109/TASL.2008.2011527.

[31] A. Srivastava, S. Bhardwaj y S. Saraswat. «SCRUM model for agile methodology». En: *2017 International Conference on Computing, Communication and Automation (ICCCA)*. 2017 International Conference on Computing, Communication and Automation (ICCCA). Mayo de 2017, págs. 864-869. DOI: 10.1109/CCAA.2017.8229928.

[32] M. Tanti. «Exploitation of "Big Data": The experience feedback of the french military health service on sanitary data». En: *2015 6th International Conference on Information Systems and Economic Intelligence (SIIE)*. 2015 6th International Conference on Information Systems and Economic Intelligence (SIIE). Feb. de 2015, págs. 1-4. DOI: 10.1109/ISEI.2015.7358716.

[33] R. Tardío, A. Mate y J. Trujillo. «An iterative methodology for big data management, analysis and visualization». En: *2015 IEEE International Conference on Big Data (Big Data)*. 2015 IEEE International Conference on Big Data (Big Data). Oct. de 2015, págs. 545-550. DOI: 10.1109/BigData.2015.7363798.

[34] F. Zhang y col. «Improvement of Pearson similarity coefficient based on item frequency». En: *2017 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR)*. 2017 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR). Jul. de 2017, págs. 248-253. DOI: 10.1109/ICWAPR.2017.8076697.

A. Dataset de prueba Cruz Roja

Cruz Roja ha proporcionado un dataset dividido en tres consultas a sus bases de datos que están descritos en las siguientes tablas.

Tabla 1: Dataset de Cruz Roja: Primera consulta

Dataset Cruz Roja 1			
Campo	Tipo	Diferentes	Completitud
PER_CODIGO	Numérico	100 %	100 %
TIDPE_PER_CODIGO_TECNICO	Numérico	3 %	100 %
TIDPE_TIPER_CODIGO_TECNICO	Numérico	2	100 %
TIDPE_FALTA_TECNICO	Nominal	1.856	100 %
FECHA	Nominal	1.204	100 %
ECO_INGRESOS	Numérico	6	100 %
ECO_TRABAJO	Nominal	2	100 %
ECO_PREST_DESEMPLEO	Nominal	1.664	100 %
ECO_PENSION	Nominal	331	100 %
ECO_PENSION_VIUDEDAD	Nominal	2	100 %
ECO_PENSION_INVALIDEZ	Nominal	2	100 %
ECO_PENSION_NO_CONT	Nominal	2	100 %
ECO_RENTA_MINIMA	Nominal	2	100 %
ECO_REDES_PERSONALES	Nominal	2	100 %
ECO_PLAN_PENSION	Nominal	2	100 %
ECO_OTROS_INGRESOS	String	192	3 %
ECO_SIN_INGRESOS	Nominal	2	100 %
ECO_NO_DEUDAS	Nominal	2	100 %
ECO_IMPAGO_HIPOTECA	Nominal	2	100 %
ECO_DEUDA_ALQUILER	Nominal	2	100 %
ECO_DEUDA_COMUNIDAD	Nominal	2	100 %
ECO_DEUDA_REC_SERV_BASICOS	Nominal	2	100 %
ECO_DEUDA_CON_ADMINISTRACION	Nominal	2	100 %
ECO_OTRAS_DEUDAS	Nominal	2	100 %
ECO_DEUDAS_PAIS_ORIGEN	Nominal	2	100 %
ECO_GASTOS_VIVIENDA	String	725	46 %
ECO_GASTOS_MANUTENCION	String	102	19 %
ECO_GASTOS_SERVICIOS	String	268	70 %
ECO_GASTOS_TELEFONO	String	100	13 %
ECO_GASTOS_MEDICACION	String	69	5 %
ECO_GASTOS_ESCOLARIZACION	String	97	3 %
ECO_GASTOS_REMESAS	String	29	2 %
ECO_GASTOS_TRANSPORTE	String	88	5 %
ECO_AYUDA_PRESTACION_HIJO	Nominal	2	100 %
ECO_AYUDAP_IMPORTE	String	227	8 %
ECO_AYUDAP_PER_REFERENCIA	Nominal	36	0 %
ECO_AYUDA_ALQUILER	Nominal	2	100 %
ECO_AYUDAA_IMPORTE	String	71	1 %
ECO_AYUDA_EMERGENCIA	Nominal	2	100 %
ECO_AYUDAE_IMPORTE	String	94	1 %

Continuación de la tabla 1			
Campo	Tipo	Diferentes	Completitud
ECO_AYUDA_MANUTENCION	Nominal	2	100 %
ECO_AYUDAM_IMPORTE	String	72	1 %
ECO_COBERT_BASICAS	Nominal	4	100 %
LAB_TRABAJA	Nominal	2	100 %
LAB_TRABAJA_TIPO	Numérico	4	100 %
LAB_TRAB_JORNADA	Numérico	3	100 %
LAB_TRAB_CONTRATO	Numérico	3	100 %
LAB_DESEMPLEO	Nominal	2	100 %
LAB_DESEM_TIEMPO	Numérico	4	100 %
LAB_TODOS_DESEMPLEADOS	Nominal	2	100 %
LAB_OTRAS_SITUACIONES	Numérico	7	100 %
LAB_MOTIVACION_EMPLEO	Numérico	4	100 %
LAB_MOTIV_EMPLEO_SITU	Numérico	4	100 %
LAB_PERS_EDAD_TRABAJAR	Numérico	10	36 %
LAB_PERS_TRABAJA_COMPLETA	Numérico	5	15 %
LAB_PERS_TRABAJA_INCOMPLETA	Numérico	4	14 %
LAB_PUESTO_TRABAJO	String	151	3 %
LAB_TRAYECTORIA_LABORAL	Nominal	2	100 %
LAB_TRAYEC_POCO_ADAPTADA	Nominal	2	100 %
LAB_TRAYEC_LAB_SITU	Numérico	4	100 %
LAB_TRAYEC_PARA_PUESTO	Numérico	4	100 %
LAB_TRAYEC_PP_SITU	Numérico	4	100 %
LAB_INCAPACIDAD_LABORAL	Nominal	2	100 %
LAB_INCAPACIDAD_LAB_TIPO	Numérico	3	100 %
LAB_INCAPACIDAD_SITU	Numérico	4	100 %
LAB_TITULACION_SITU	Numérico	4	100 %
FAM_VIVEN_DOMICILIO	Numérico	10	75 %
FAM_VIVEN_PAREJA	Nominal	2	100 %
FAM_VIVEN_DESCENDIENTES	Nominal	2	100 %
FAM_VIVEN_DESC_MENOS_3	Numérico	8	25 %
FAM_VIVEN_DESC_MAS_3	Numérico	11	20 %
FAM_VIVEN_ASCENDENTES	Numérico	5	3 %
FAM_VIVEN_OTROS	Numérico	12	6 %
FAM_VIVEN_FAM_DEPENDIENTES	Numérico	8	1 %
FAM_HORAS_DEDICACION	Numérico	6	100 %
FAM_DIAS_DEDICACION_DEPEND	Numérico	24	7 %
FAM_SERVICIOS	Numérico	8	100 %
FAM_TIPO_FAMILIA	Numérico	11	100 %
FAM_BAJO_RESPONSABILIDAD	Nominal	4	100 %
FAM_PRIV_HIPOTECA	Nominal	2	100 %
FAM_TEMP_VIVIENDA	Nominal	2	100 %
FAM_VACACIONES	Nominal	2	100 %
FAM_COMIDA	Nominal	2	100 %
FAM_GASTOS_IMPREVISTOS	Nominal	2	100 %
FAM_DISPO_TELEFONO	Nominal	2	100 %
FAM_DISPO_TV_COLOR	Nominal	2	100 %

Continuación de la tabla 1			
Campo	Tipo	Diferentes	Completitud
FAM_DISPO_LAVADORA	Nominal	2	100 %
FAM_DISPO_COCHE	Nominal	2	100 %
FAM_PROBLEMA_EXISTENCIA	Nominal	2	100 %
FAM_BUEN_ENTORNO	Nominal	2	100 %
FAM_PROBLEMA_DIFICULTAD	Nominal	2	100 %
FAM_APOYO_ENTORNO	Nominal	2	100 %
FAM_PROBLEMA_SITU	Nominal	2	100 %
FAM_CONFLICTO_FAMILIAR	Nominal	2	100 %
FAM_CONFLICTO_FAMILIAR_TIPO	Nominal	4	100 %
FAM_PROB_DIAGNOSTICADOS	Nominal	2	100 %
FAM_PROB_DEPENDENCIAS	Nominal	2	100 %
FAM_PROB_DISCAPACIDAD	Nominal	2	100 %
FAM_PROB_SALUD_MENTAL	Nominal	2	100 %
FAM_PROB_MENORES_INST	Nominal	2	100 %
FAM_PROB_MUJ_MALTRATADAS	Nominal	2	100 %
FAM_PROB_ABUSO_SEXUAL	Nominal	2	100 %
FAM_PROB_SIN_HOGAR	Nominal	2	100 %
FAM_PROB_ENFERMEDAD	Nominal	2	100 %
FAM_PROB_OTROS	Nominal	2	100 %
FAM_CONDUCTA_DELICTIVA	Nominal	2	100 %
FAM_COND_RECLUSOS	Nominal	2	100 %
FAM_COND_MENORES_INF	Nominal	2	100 %
FAM_COND_OTROS	Nominal	2	100 %
FAM_NECESIDAD_APOYO	Nominal	2	100 %
FAM_PETICION_ANT_APOYO	Nominal	2	100 %
FAM_AYUDA_NO_RECIBIDA	Nominal	2	100 %
FAM_AYUDA_SI_RECIBIDA	Nominal	2	100 %
FAM_SOLUCIONO_PROBLEMA	Nominal	2	100 %
FAM_APOYO_CRE	Nominal	2	100 %
AMB_SITU_VIVIENDA	Nominal	13	100 %
AMB_VIVIENDA_ADAPTADA	Nominal	2	100 %
AMB_TIPO_VIVIENDA	Nominal	4	100 %
AMB_CONDI_HABITABILIDAD	Nominal	6	100 %
AMB_EQUI_ADECUADO	Nominal	2	100 %
AMB_EQUI_INADECUADO	Nominal	2	100 %
AMB_EQUI_SIN_ASEO	Nominal	2	100 %
AMB_EQUI_SIN_LUZ	Nominal	2	100 %
AMB_EQUI_SIN_AGUA	Nominal	2	100 %
AMB_EQUI_SIN_CALEFACCION	Nominal	2	100 %
AMB_EQUI_SIN_ELECTRODOMESTICOS	Nominal	2	100 %
AMB_EQUI_SIN_OTRO	Nominal	2	100 %
AMB_TIPO_ENTORNO	Nominal	6	100 %
AMB_CARACT_ENTORNO	Numérico	5	100 %
AMB_LUGAR_RESIDE	Numérico	4	100 %
AMB_ACCESIBILIDAD_RECURSOS	Nominal	5	100 %
AMB_ACCESO_RECURSOS	Nominal	4	100 %

Continuación de la tabla 1			
Campo	Tipo	Diferentes	Completitud
AMB_CONDI_ALOJAMIENTO	Nominal	4	100 %
AMB_ACCESIBILIDAD	Nominal	4	100 %
SOC_REDES_APOYO	Nominal	2	100 %
SOC_REDES_APOYO_PERS	Numérico	4	100 %
SOC_ACEP_DERECHOS_SOCIALES	Nominal	2	100 %
SOC_DISCRIMINACION	Nominal	2	100 %
SOC_DISCRIMINACION_SEXO	Nominal	2	100 %
SOC_DISCRIMINACION_EDAD	Nominal	2	100 %
SOC_DISCRIMINACION_ORIGEN	Nominal	2	100 %
SOC_RACISMO	Nominal	2	100 %
SOC_PRACTICA_PERSONAL	Nominal	2	100 %
SOC_ACOSO	Nominal	2	100 %
SOC_PERTENECIA_GRUPO_ETNICO	Numérico	4	100 %
SOC_SEXO	Numérico	4	100 %
SOC_EDAD	Numérico	4	100 %
SOC_PRACTICAS_RELIGIOSAS	Nominal	4	100 %
SOC_NIVEL_ESTUDIOS	Numérico	1	100 %
SOC_HOMOLOGACION_ESTUDIOS	Nominal	2	100 %
SOC_FRACASO_ESCOLAR	Nominal	2	100 %
SOC_AISLAMIENTO_SOCIAL	Nominal	2	100 %
SOC_AISLA_LENGUA	Nominal	2	100 %
SOC_AISLA_CULTURA	Nominal	2	100 %
SOC_AISLA_ETNIA	Nominal	2	100 %
SOC_AISLA_INTERNAMIENTO	Nominal	2	100 %
SOC_PARTICIPACION_SOCIAL	Nominal	2	100 %
SOC_PARTIC_FORMAL	Numérico	3	100 %
SOC_ANTECEDENTES_PENALES	Nominal	2	100 %
SOC_ANTEC_PENALES_TIPO	Numérico	4	100 %
SOC_ANTEC_PENALES_SITU	Numérico	4	100 %
SOC_ORIENTACION_MEDIO	Nominal	2	100 %
SOC_ORIENTACION_MEDIO_DET	Nominal	29	0 %
SOC_ORIENTACION_MEDIO_SITU	Numérico	4	100 %
SAL_DISPONE_TARJETA	Nominal	2	100 %
SAL_PROBLEMAS_SALUD	Nominal	2	100 %
SAL_RECOMEN_PROD_APOYO	Nominal	2	100 %
SAL_TRATAMIENTO_NECESARIO	Nominal	2	100 %
SAL_PROB_ACCESO_MED	Nominal	2	100 %
SAL_NO_ATENDIDO	Nominal	2	100 %
SAL_ALIMENTACION_VARIADA	Nominal	2	100 %
SAL_EJERCICIO_FISICO	Nominal	2	100 %
SAL_HABITO_SALUDABLE	Numérico	4	100 %
SAL_DISCAPACIDAD	Nominal	2	100 %
SAL_DISCAPACIDAD_TIPO	Numérico	4	100 %
SAL_DISCAPACIDAD_SITU	Numérico	4	100 %
SAL_DEPENDENCIA	Nominal	2	100 %
SAL_DEPENDENCIA_TIPO	Numérico	4	100 %

Continuación de la tabla 1			
Campo	Tipo	Diferentes	Completitud
SAL_ENFERMEDAD_FISICA	Nominal	2	100 %
SAL_ENFER_FISICA_DIABETES	Nominal	2	100 %
SAL_ENFER_FISICA_VIH	Nominal	2	100 %
SAL_ENFER_FISICA_CANCER	Nominal	2	100 %
SAL_ENFER_FISICA_TUBERCULOSIS	Nominal	2	100 %
SAL_ENFER_FISICA_HEPATITIS	Nominal	2	100 %
SAL_ENFER_FISICA_OTRA	Nominal	2	100 %
SAL_ENFER_MENTALES	Nominal	2	100 %
SAL_ENFER_MENT_DEPRESION	Nominal	2	100 %
SAL_ENFER_MENT_ALZHEIMER	Nominal	2	100 %
SAL_ENFER_MENT_OTRA	Nominal	2	100 %
SAL_ESTADO_EMOCIONAL	Nominal	2	100 %
SAL_ESTADO_EMOCIONAL_VAL	Numérico	6	19 %
SAL_ADICCIONES	Nominal	2	100 %
SAL_ADICCIONES_ALCOHOL	Nominal	2	100 %
SAL_ADICCIONES_DROGAS	Nominal	2	100 %
SAL_ADICCIONES_SITU	Numérico	4	100 %
SAL_DIAGNOSTICADA	Nominal	2	100 %
SAL_PETICION_AYUDA	Nominal	2	100 %
SAL_PETICION_AYUDA_TIPO	Numérico	3	100 %
SAL_SOLUCION_PROBLEMA	Nominal	2	100 %
SAL_SOL_PROBLEMA_SEGUIMIENTO	Nominal	2	100 %
SAL_BIENESTAR_SITU	Numérico	4	100 %
PER_NACIONAL	Nominal	2	100 %
PER_NACIONAL_TIPO	Nominal	2	100 %
PER_NO_NACIONAL	Nominal	2	100 %
PER_SITUA_ADMINISTRATIVA	Numérico	8	100 %
PER_SITUA_ADMINISTRATIVA_SITU	Numérico	4	100 %
PER_FECHA_PER_RESIDENCIA	Nominal	388	2 %
PER_FECHA_PER_TRABAJO	Nominal	1.114	8 %
PER_CARNET_CONDUCIR	Nominal	2	100 %
PER_CARNET_CONDUCIR_TIPO	Numérico	3	100 %
PER_DISPONIBILIDAD	Nominal	2	100 %
PER_DISPO_HORARIA	Nominal	2	100 %
PER_DISPO_HORARIA_TIPO	Numérico	4	100 %
PER_DISPO_GEOGRAFICA	Nominal	2	100 %
PER_DISPO_GEOGRAFICA_TIPO	Numérico	4	100 %
PER_DISPONIBILIDAD_SITU	Numérico	4	100 %
PER_CONOCIMIENTO_IDIOMA	Numérico	4	100 %
PER_CUIDAD_IMAGEN	Numérico	4	100 %
PER_LOCUS_CONTROL	Numérico	4	100 %
PER_VIOLENCIA	Numérico	4	100 %
PER_VIOLENCIA_FAMILIAR	Nominal	2	100 %
PER_VIOLENCIA_FAMILIAR_SITU	Numérico	4	100 %
CENTR_CODIGO_ATENCION	Numérico	584	68 %
CODIGO_ACTUACION_RI	String	0	0 %

Continuación de la tabla 1			
Campo	Tipo	Diferentes	Completitud
RESULTADO_VALORACION_RI	String	0	0 %
FAM_VIVEN_HERMANOS	Numérico	6	1 %
Final de la tabla 1			

Campo	Tipo	Diferentes	Completitud
SECUENCIA_PPI	Numérico	9.735	100 %
CENTR_CODIGO_PPI	Numérico	52	100 %
TIDPE_FALTA_USU	Nominal	2.848	100 %
TIDPE_TIPER_CODIGO_USU	Numérico	100 %	100 %
TIDPE_PER_CODIGO_USU	Numérico	100 %	100 %
TIDPE_FALTA_TEC	Nominal	1.497	100 %
TIDPE_TIPER_CODIGO_TEC	Numérico	100 %	100 %
TIDPE_PER_CODIGO_TEC	Numérico	100 %	100 %
ESTADO	Nominal	3	100 %
TIDPE_FALTA_TUTOR	Nominal	1.167	87 %
TIDPE_TIPER_CODIGO_TUTOR	Numérico	3	87 %
TIDPE_PER_CODIGO_TUTOR	Nominal	1.555	87 %
ACCION	Nominal	12	100 %
CLASE	Nominal	2	100 %
DSACCION	Nominal	12	100 %
TIPO_ACCION	Numeric	86	100 %
DSTIPO_ACCION	Nominal	53	100 %
SUBTIPO_ACCION	Numeric	216	100 %
DSSUBTIPO_ACCION	Nominal	184	60 %
CENTRO_CODIGO_ACT	Nominal	500	100 %
PROYECTO_TACTOFI	Nominal	190	100 %
DSCENTRO	Nominal	536	100 %
DSPROYECTO	Nominal	168	100 %
PROGRAMA_TACTOFI	Nominal	31	100 %
PROGRAMA	Nominal	31	100 %
FC_ELABORACION	Nominal	376	100 %
FECHA_CREACION	Nominal	272	100 %
SECUENCIA	Numérico	100 %	100 %
CAB_ACTIV_SECUENCIA	Numérico	100 %	100 %
CODIGO_PLAN	Nominal	6	100 %
DSPLAN	Nominal	6	100 %
PERIODICIDAD	Nominal	1.297	100 %
TACTOFI	Nominal	3.526	100 %
CENTRO_CODIGO	Nominal	536	100 %
ASIGNADO	Nominal	2	100 %
COD_GRUPO	Nominal	854	100 %
TACT_SECUENCIA	Numérico	177	100 %
CODIGO_PROYECTO	Nominal	173	100 %
CODIGO_PROGRAMA	Nominal	31	100 %

Tabla 2: Dataset de Cruz Roja: Segunda consulta

Campo	Tipo	Diferentes	Completitud
SECUENCIA	Numérico	100 %	100 %
ACTIVIDAD_REAL_SECUENCIA	Numérico	77 %	100 %
TSUBPROYECT_PACK_SECUENCIA	Numérico	18	100 %
TIPO	Nominal	2	100 %
FECHA_INICIO	Nominal	2.980	100 %
FECHA_FIN	Nominal	2.961	100 %
HORA1_INICIO	Nominal	1.664	98 %
HORA1_FIN	Nominal	1.664	98 %
HORA2_INICIO	Nominal	331	0 %
HORA2_FIN	Nominal	331	0 %
HORA3_INICIO	Nominal	258	0 %
HORA3_FIN	Nominal	258	0 %
TIDPE_PER_CODIGO	Numérico	100 %	100 %
TIDPE_TIPER_CODIGO	Numérico	100 %	100 %
TIDPE_FALTA	Nominal	2.881	100 %
GRUPO_CODIGO	String	0	0 %
LUNES	Nominal	2	100 %
MARTES	Nominal	2	100 %
MIERCOLES	Nominal	2	100 %
JUEVES	Nominal	2	100 %
VIERNES	Nominal	2	100 %
SABADO	Nominal	2	100 %
DOMINGO	Nominal	2	100 %
PORC_PAGO	String	0	0 %
PORC_DESC_CR	String	0	0 %
PAGO_ALTA	String	0	0 %
PAGO_DEPOSITO	String	0	0 %
ASISTENCIA	Nominal	7	99 %

Tabla 3: Dataset de Cruz Roja: Tercera consulta